THE UNIVERSITY
of EDINBURGH

Combining Genome-Wide Association Studies, Polygenic Risk Scores and SNP-SNP interactions to investigate the genomic architecture of human complex diseases: More than the sum of its parts

Joeri Jeroen Meijsen

THE UNIVERSITY *of* EDINBURGH

A thesis presented for the degree of
Doctor of Philosophy

Centre for Genomic and Experimental Medicine
Institute of Genetics and Molecular Medicine
The University of Edinburgh
Edinburgh
Scotland
2018

Declaration

I declare that this thesis has been composed by me, that the work described in this thesis is my own, except where otherwise stated, and that the work described in this thesis has not been submitted for any other degree or professional qualification.

Chapters 2 and 3 both are published as multi-author papers. These papers have been written in their entirety by me, co-authors have seen the final draft before submission and made comments/suggestions related to their respected fields of expertise.

_____

Joeri Meijsen

Acknowledgements

First of all, I would like to thank my supervisor Kristin Nicodemus, for her support, advice and especially her almost unlimited patience over the past 3 years. I have been very lucky to be supervised by such an expert in the fields of statistical genetics, machine learning and psychiatry. I learned a great deal from her and I hope that we will be able to collaborate on many projects in the future. I would also like to thank Riccardo Marioni and Ian Deary for being my other supervisors and for sharing their extensive knowledge of all things related to cognition and psychology. I would like to thank Pippa Thompson and Andrew McIntosh for their in-depth and helpful suggestions as my external members during my thesis committee meetings and David Porteous for always giving me helpful advice. Last but not least I would like to thank John Ireland, who never once got angry when I bugged him numerous times a day with cluster computer issues.

Oscar Lao Grueso was my supervisor for both my MSc internships at the Erasmus University Medical Center Rotterdam. Oscar helped me to get the required qualifications and expertise to be even considered for this position. Without his encouragement and help I would have never gotten into this PhD and for that I would like to thank him.

I would like to thank all Nicodemus lab group members and students for making my PhD an enjoyable experience, thank you: Elvina, Siân, Ksenia, Alex, Thalia, Lara and Linda. Thanks to my friends in the Edinburgh University Judo Club for making me feel old on a daily basis. Finally, I would like to thank my parents and "oma" for their everlasting support and encouragement throughout my entire educational journey, even though 3 years ago they would have loved me utter the following sentence: "mum…dad…I am going to do a PhD in Delft".

Abstract

Major Depressive Disorder is a devastating psychiatric illness with a complex genetic and environmental component that affects 10% of the UK population. Previous studies have shown that that individuals with depression show poorer performance on measures of cognitive domains such as memory, attention, language and executive functioning. A major risk factor for depression is a higher level of neuroticism, which has been shown to be associated with depression throughout life. Understanding cognitive performance in depression and neuroticism could lead to a better understanding of the aetiology of depression. The first aim of this thesis focused on assessing phenotypic and genetic differences in cognitive performance between healthy controls and depressed individuals and also between single episode and recurrent depression. A second aim was determining the capability of two decision-tree based methods to detect simulated gene-gene interactions. The third aim was to develop a novel statistical methodology for simultaneously analysing single SNP, additive and interacting genetic components associated with neuroticism using machine leaning.

To assess the phenotypic and genetic differences in depression, 7,012 unrelated Generation Scotland participants (of which 1,042 were clinically diagnosed with depression) were analysed. Significant differences in cognitive performance were observed in two domains: processing speed and vocabulary. Individuals with recurrent depression showed lower processing speed scores compared to both controls and individuals with single episode depression. Higher vocabulary scores were observed in depressed individuals compared to controls and in individuals with recurrent depression compared to controls. These significant differences could not be tied to significant single locus associations. Derived polygenic scores using the large CHARGE processing speed GWAS explained up to 1% of variation in processing speed performance among individuals with single episode and recurrent depression.

Two greedy non-parametric decision-tree based methods – C5.0 and logic regression - were applied to simulated gene-gene interaction data from Generation Scotland. Several gene-gene interactions were simulated under multiple scenarios (e.g. size, strength of association levels and the presence of a polygenic component) to assess the power and type I error. C5.0 was found to have an increased power with a conservative type I error using simulated data. C5.0 was applied to years of education as a proxy of educational attainment in 6,765 Generation Scotland participants. Multiple interacting loci were detected that were associated with years of education, some most notably located in genes known to be associated with reading and spelling (*RCAN3*) and neurodevelopmental traits *(NPAS3*).

C5.0 was incorporated in a novel methodology called Machine-learning for Additive and Interaction Combined Analysis (MAICA). MAICA allows for a simultaneous analysis of single locus, polygenic components, and gene-gene interaction risk factors by means of a machine learning implementation. MAICA was applied on neuroticism scores in both Generation Scotland and UK Biobank. The MAICA model in Generation Scotland included 151 single loci and 11 gene-gene interaction sets, and explained ~6.5% of variation in neuroticism scores. Applying the same model to UK Biobank did not lead to a statistically significant prediction of neuroticism scores.

The results presented in this thesis showed that individuals with depression performed significantly lower on the processing speed tests but higher on vocabulary test and that 1% of variation in processing speed can be explained by using a large processing speed GWAS. Evidence was provided that C5.0 had increased power and acceptable type I error rates versus logic regression when epistatic models exist – even with a strong underlying polygenic component, and that MAICA is an efficient tool to assess single locus, polygenic and epistatic components simultaneously. MAICA is open-source, and will provide a useful tool for other researchers of complex human traits who are interested in exploring the relative contributions of these different genomic architectures.

Lay Abstract

Depression is a common psychiatric condition with both genetic and environmental risk factors. Individuals with depression have shown a lower performance in multiple cognitive domains and higher levels of neuroticism. Understanding traits associated with depression can improve understanding of depression itself. The work presented in this thesis aimed to investigate the phenotypic (observable) and genetic (non-observable) differences of cognitive performance in depression and genetic association with educational attainment and neuroticism.

We assessed cognitive performance in 7,012 Scottish individuals of which 1,042 with a depression diagnosis. Significant differences in cognitive performance were observed in processing speed and vocabulary: between healthy controls and individuals with depression but also within the group of depressed individuals. Observed differences in cognitive performance could not be linked to small genetic differences and the sum of multiple small genetic differences accounted for 1% of variation in differences in processing speed.

The genetic contribution for most diseases is thought to be attributed to small genetic differences and the sum of independently acting genetic variants. However these risk factors explain only a small percentage of variation in most diseases and traits. An overlooked genetic component is in gene-gene interactions, where multiple dependent genetic differences act together to affect the disease. Accurately detecting gene-gene interactions will potentially increase the amount of variation explained in diseases and traits. Two different methods, called C5.0 and logic regression, were applied on simulated gene-gene interaction data under different conditions to assess their capability of detecting genetic interactions in a large genetic dataset. Significant differences in accurately detecting gene-gene interactions were observed. C5.0 showed higher power (percentage of accurately detected simulated interactions) and an almost 0 type I error rate (percentage of detected interactions when none exists in the data). C5.0 was applied to detect putative interacting loci involved with the number of years

of education in 6,765 Scottish individuals. Interacting loci observed were previously found in genes involved with reading and spelling abilities *(RCAN3*) but also with neuropsychiatric conditions such as schizophrenia, bipolar disorder, depression and ADHD (*NPAS3*).

Combining small genetic differences, sum of small genetic differences and gene-gene interaction risk factors into one methodology may explain more variation in a disease or trait such as neuroticism. To examine this possibility, we developed a novel methodology called MAICA to assess all three components in two independent large cohort studies, Generation Scotland and UK Biobank. Using MAICA, we were able to explain a small percentage of variation in Generation Scotland however this did not replicate in UK Biobank.

The work presented in this thesis has shown a significant difference in cognitive performance (both positive and negative) between individuals with depression and controls but also between single episode and recurrent depression. 1% of variation in differences in processing speed performance was explained using a method that sums multiple small genetic differences. Novel evidence was provided to show the existence of gene-gene interactions however this is somewhat ambiguous. MAICA is an open-source tool to assess single locus, polygenic and gene-gene interactions simultaneously to determine the genetic contribution to a trait.

List of abbreviations

CART: Classification and Regression Trees

CHARGE: The Cohorts for Heart and Aging Research in Genomic Epidemiology

CONVERGE: China, Oxford and Virginia Commonwealth University Experimental Research on Genetic Epidemiology

CVF: Categorical Verbal Fluency

DSM-IV: Diagnostic and Statistical Manual of Mental Disorders-IV

DST: Digit Symbol substitution Task

DZ: Dizygotic

EA: Educational attainment

EF: Executive Functioning

EPQ-R: Eysenck Personality Questionnaire-Revised

GAIN: Genetic Association Information Network

GCTA: Genome-wide Complex Trait Analysis

GEC: Genetic type 1 Error Calculator

GP: General Practitioner

GPAS: Genetic Programming for Association Studies

GPC-2: Genetics of Personality Consortium

GS:SFHS: Generation Scotland: the Scottish Family Health Study

GWAS: Genome-Wide Association Study

GWEIS: Genome-Wide by Environment Interaction Study

IBD: Identical by Descent

ICD-10: International Statistical Classification of Diseases and Related Health Problems

LASSO: Least Absolute Selection and Shrinkage Operator

LD: Linkage Disequilibrium

LDAK: Linkage-Disequilibrium Adjusted Kinships

LE: Linkage Equilibrium

LM1: Logical Memory Immediate

LM2: Logical Memory Delayed

LOD: Log of the Odds Ratio

MAF: Minor Allele Frequency

MAICA: Machine-learning for Additive and Interaction Combined Analysis

MDD: Major Depressive Disorder

MDE: Major Depressive Episode

MHVS: Mill Hill Vocabulary Scale

ML: Machine learning

MZ: Monozygotic

NART: National Adult Reading Test

NHS: National Health Service

OLS: Ordinary Least Squares

OR: Odds Ratio

PGC: Psychiatric Genomics Consortium

PRS/PGRS: Polygenic Score / PolyGenic Risk Score

PVF: Phonemic Verbal Fluency

QC: Quality Control

RF: Random Forrest

SA: Simulated Annealing

SCID: Structured Clinical Interview for DSM-IV

SES: Socio Economic Status

SNP: Single Nucleotide Polymorphism

SSRI: Selective Serotonin Reuptake Inhibitors

SWM: Spatial Working Memory

UK-B: United Kingdom Biobank

VFT: Verbal Fluency Total

WAIS: Wechsler Adult Intelligence Scale

WHO: World Health Organisation

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1 Introduction

1.1 Literature review

1.1.1 Symptoms and epidemiology of Major Depressive Disorder

Major Depressive Disorder (MDD) is a common mental disorder having a prevalence of one in seven to one in ten in the United Kingdom (D. J. Smith *et al.*, 2013).The World Health Organisation (WHO) has ranked MDD the 4th leading cause of disability worldwide (Murray *et al.*, 1996). The WHO has also projected that by 2020 MDD will rise to be the 2nd leading cause of disability worldwide (Lopez and Murray, 1998).

In 2000, adult MDD cost the United Kingdom approximately £9 billion, with an estimated 110 million working days lost (Thomas and Morris, 2003). In the USA in the same year, MDD accounted for costs totalling over $80 billion for treatment and disability (Greenberg *et al.*, 2003). As both groups (Greenberg *et al.*, 2003) published their results in 2003 and MDD prevalence has increased, so has the cost. These numbers indicate that besides individual suffering, MDD has a global economic impact.

MDD is characterised by having one or multiple Major Depressive Episodes (MDE). Symptoms of a (MDE) can be highly variable between individuals. A MDE is mainly a constant and persistent feeling of sadness, sleep disturbances, low mood, guilt, hopelessness and loss of interest in activities that previously gave the individual pleasure, among others.

Worldwide, two operational diagnostic systems are used to classify MDD using symptom-based criteria. However the Diagnostic and Statistical Manual of Mental Disorders-IV (DSM-IV) of the American Psychiatric Association (American

Psychiatric Association, 1994) and the International Statistical Classification of Diseases and Related Health Problems (ICD-10) of the World Health Organisation (World Health Organisation, 2004) show discrepancies in these criteria symptoms in making a diagnosis. According to the DSM-IV, the diagnosis of MDD includes: A minimum of five of the following nine symptoms, present nearly every day during the last two weeks or more including at least point 1 or 2 (Table 1.1).

1.	Depressed mood or irritable most of the day, as indicated by either subjective report or observation made by others.
2.	Decreased interest or pleasure in most activities
3.	Significant weight change (5%) or change in appetite
4.	Change in sleep: Insomnia or hypersomnia
5.	Change in activity: Psychomotor agitation or retardation
6.	Fatigue or loss of energy
7.	Guilt/worthlessness: Feelings of worthlessness or excessive or inappropriate guilt
8.	Concentration: diminished ability to think or concentrate, or increased indecisiveness
9.	Suicidality: Thoughts of death or suicide, or has suicide plan

Table 1.1: DSM-IV symptoms of MDD

On the other hand, the ICD-10 diagnosis of MDD includes three typical depressive symptoms (depressed mood, anhedonia, and reduced energy), two of which should be present to determine depressive disorder diagnosis (Table 1.2). Other common symptoms are:

A.	Reduced concentration and attention
B.	Reduced self-esteem and self confidence
C.	Ideas of guilt and unworthiness
D.	Bleak and pessimistic views of the future
E.	Ideas or acts of self-harm or suicide
F.	Disturbed sleep
G.	Diminished appetite

Table 1.2: ICD-10 symptoms of MDD

The average age of onset of MDD was estimated to be 26 years (Zisook *et al.*, 2007). However around 40% of individuals had their first MDE between the age of 10 and 21 (Figure 1.1) which are also the puberty and post-puberty years. Due to this skewed distribution a median age of onset would be a more accurate representation making the age of onset of MDD around 18 years.

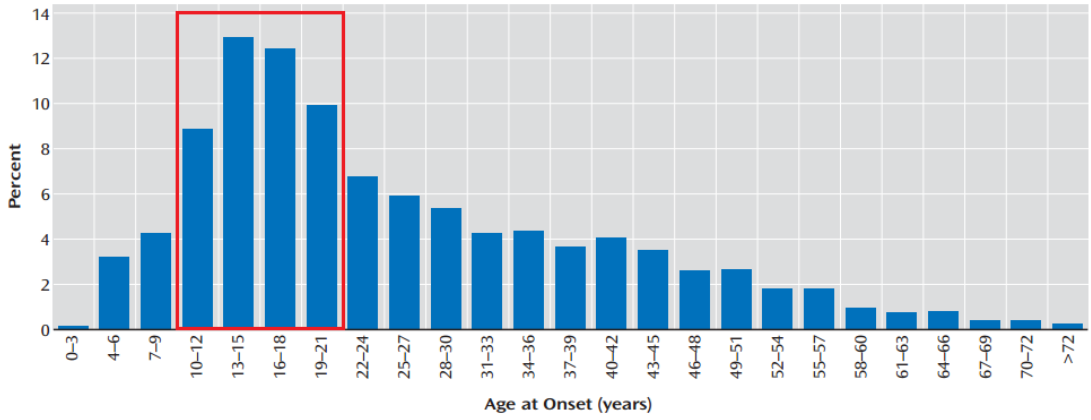


Figure 1.1: Distribution of Age at Onset of first MDE

The red square represents 40% of cases. Image adapted from (Zisook et al., 2007).

Approximately 50% or more of MDD cases will have a second depressive episode within six months after the initial index episode (American Psychiatric Association, 1994) and 15% will have MDD as a lifelong chronic illness (Eaton *et al.*, 2008), making MDD a highly recurrent disorder.

Large population-based studies have shown that women are twice as likely to be diagnosed with MDD compared to men (Kessler *et al.*, 1993; Kuehner, 2003). However, studies have hypothesised that this might be inflated due to the larger social stigma MDD has for men, making them less likely to see a general practitioner (GP) compared to women (Oliffe and Phillips, 2008; Oliffe *et al.*, 2016).

The WHO World Mental Health (WMH) survey initiative studied cross-national differences in MDD prevalence by performing surveys using diagnostic interviews

based on the DSM-IV (Kessler and Bromet, 2013). The 18 study countries were divided by income into a high income and low-middle income group. Kessler and Bromet (2013) found no evidence of substantial cross-national differences in MDD prevalence. The highest prevalence estimates of lifetime MDD were observed in high income countries, i.e., France and the United States, which might be explained by the larger extent of income inequality and larger differences in socio-economic status (Kessler and Bromet, 2013).

1.1.2 Correlations between MDD and health/lifestyle factors

Here, I describe health/lifestyle factors most often observed being associated with MDD. Most of these factors will be used in subsequent analyses.

Lyall *et al.* (2016) performed a large population wide study (*n*=172,751) looking into the characteristics of MDD, and observed a trend between smoking and MDD (single episode, recurrent-mild and recurrent-severe MDD) showing that not only the prevalence of smoking in depressed individuals is higher than non-depressed individuals but also the more severe the MDD case the more likely individuals are to smoke.

In addition, Lyall *et al.* (2016) observed no difference in alcohol consumption between controls and MDD subtypes with around 20% reporting daily or almost daily alcohol use. They hypothesised that a larger proportion of individuals with (severe) depression stopped drinking alcohol for health and medical reasons, such as taking antidepressants.

The combination of diabetes and MDD occurs around two times more often than one would see by chance (Anderson, Freedland and Lustman, 2001). A large meta-analysis (*n*=50,000) of type 2 diabetes (acquired diabetes, most often a result of

obesity) has shown that the incidence of MDD is 24% higher than in controls (Nouwen *et al.*, 2010). The association with diabetes and MDD is unclear, with some studies claiming a bi-directional effect (Knol *et al.*, 2006; Golden *et al.*, 2008; Mezuk *et al.*, 2008). Knol *et al.*, 2006 meta-analysed nine cohort studies and reported a 37% increased risk of type 2 diabetes after taking out the effect of covariates such as sex, body mass index and poverty. However, Knol *et al.*, (2006) observed substantial heterogeneity across the nine studies with relative risks varying between 1.03 (non-significant) and 2.50.

Finally, as expected, antidepressant and mood stabiliser medication (psychotropic medication) usage was associated with MDD severity. However the medications prescribed vary in dosage, brands and target different mechanisms in the human body. It is worth noting that non-MDD individuals may report using psychotropic medication for non-psychiatric indications such as chronic pain relief (Lyall *et al.*, 2016).

1.2 Overview of cognitive domains

1.2.1 Introduction to intelligence

The question "What is intelligence?" is almost as old as the field of psychology itself and researchers have thus far not been able to agree on a definition. A clear and easy to understand definition was given by (Gottfredson, 1997)

"Intelligence is a very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly, and learn from experience. It is not merely book-learning, a narrow academic skill, or test-talking smarts. Rather, it reflects a broader and deeper capability for comprehending our surroundings, 'catching on', 'making sense' of things, or 'figuring out' what to do"

1.2.2 Structure of general intelligence/cognition

Variation in intelligence can be divided into three levels (Figure 1.2) (Deary, 2012). Level 1, people differ in their performance of specific narrow tests that assess various cognitive domains (test specific variation). Scores of all narrow tests of cognitive domains correlate positively (Deary, Penke and Johnson, 2010); however, some level 1 tests correlate more strongly, leading to more subsets of level 1 tests. It is found that these subsets of level 1 tests all measure the same broad cognitive domain, thus a latent trait can be extracted to represent the common variance at the domain level (level 2: domain specific variation). At level 3, individuals who perform well in a certain cognitive domain also tend to perform well in other cognitive domains (Deary, 2014). This can be extracted into a trait that represents variance of a general intelligence/cognition called 'g'.

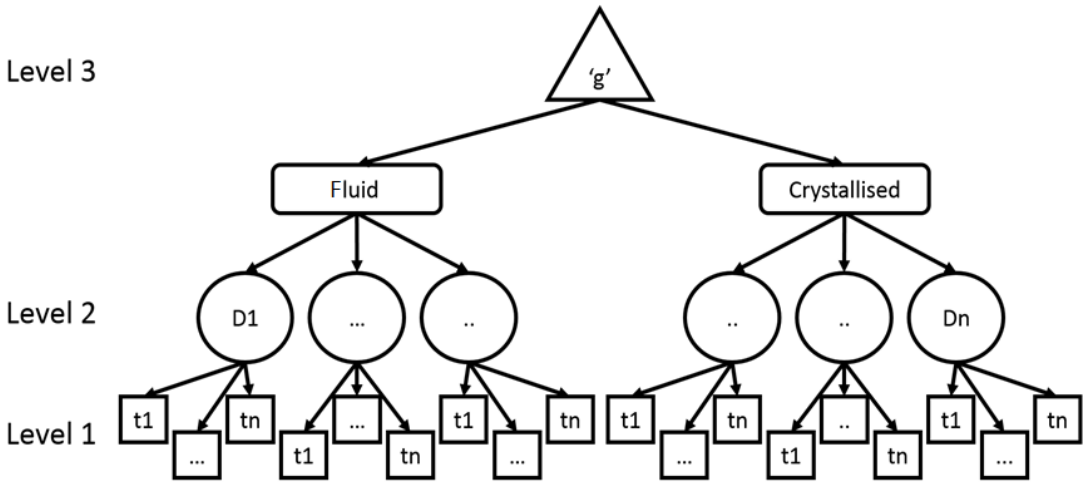


Figure 1.2: A hierarchical structure of variance in intelligence.

Level 1 (squares) indicate variance specific to tests assessing one or multiple cognitive domains. Level 2 (circles) indicates variance within a certain cognitive domain. Although not a separate level of multiple cognitive domains can be grouped together under two separate banners namely 'fluid' and 'crystallised' performance (rectangles). Finally, individuals who perform well on one cognitive domain perform well on all cognitive domains leading to level 3 (triangle) of variance general intelligence/cognitive performance (g).

1.2.3 Fluid vs crystallised cognitive ability

One might say there is a fourth level of variation within intelligence. Level 2 broad cognitive domains can be grouped into two separate subgroups, namely, 'fluid' and 'crystallised' cognitive ability (Figure 1.2).

Fluid intelligence involves the capacity to reason and solve novel problems without any pre-existing knowledge gained by previous experience (Jaeggi *et al.*, 2008). Cognitive domains falling in the fluid category are: Executive Function/Reasoning, while not exactly the same, executive functioning and reasoning are often used interchangeably. More accurately would be to say that executive functioning is an umbrella term for multiple cognitive processes that are necessary for the control of goal-setting behaviours. This includes basic cognitive processes such as inhibition, inhibitory control, working memory, and cognitive flexibility but also higher order executive functioning functions such as reasoning and problem solving (Chan *et al.*, 2008; Diamond, 2014). Processing Speed/Attention focusses on the ability to concentrate on a selective part of information, while ignoring other perceivable and sometimes distracting information (Salthouse, 1996). Memory refers to the ability to store, stabilise and retrieve information. This can either be immediate/short term memory or delayed/long term memory (Baddeley, 2007). The final domain is Spatial Ability which involves navigation, estimating distance between objects and to understand, reason and subsequently remember spatial relations among objects. This can be either in 2D, 3D and even 4D (Carroll, 1993).

Crystallised intelligence involves the ability to use skills, knowledge, and experience to find a solution to a given task. Language falls in this section and can be measured using oral or written responses or the comprehension of the previous two. Measuring language is way to measure general knowledge and is mostly measured by vocabulary (Ritchie, 2015).

There is a clear difference between fluid and crystallised intelligence when looking at performance over a lifetime. Both fluid and crystallised intelligence performance increase from birth to roughly the mid-thirties, where crystallised intelligence shows a steeper increase. After the mid-thirties fluid intelligence decreases over time (Figure 1.3) whereas crystallised intelligence remains relatively stable into old age (Tucker-Drob, 2009).

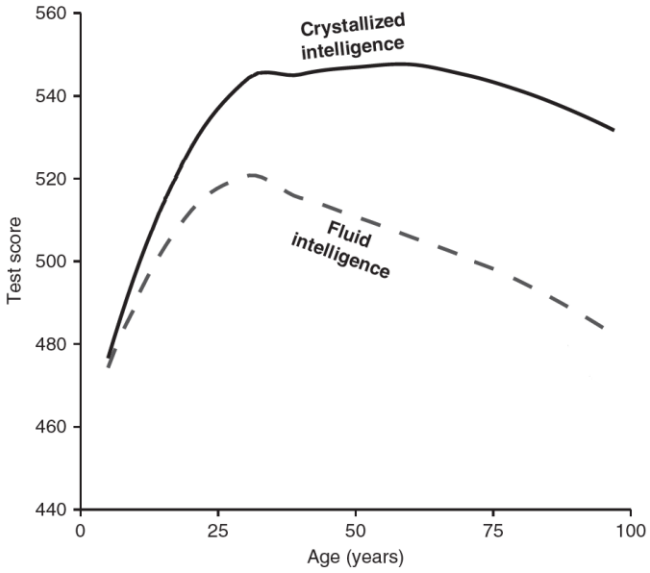


Figure 1.3: Cognitive ability of 6000 individuals over time (age).

Dotted line is fluid intelligence and solid line is crystallised intelligence. Image taken from Tucker-Drob, 2009.

1.2.4 Cognitive ability measures

Here I describe in detail cognitive ability measures used in subsequent analyses and others in shorter detail.

Executive Functioning/reasoning can be measured by a wide range of tests. Verbal reasoning and (non-verbal) matrix reasoning are however most widely used. Executive functioning is measured by someone's verbal fluency. Verbal fluency can be subdivided in two main categories namely Categorical Verbal Fluency (CVF) and

Phonemic Verbal Fluency (PVF). For the CVF test participants are asked to name as many words from a specific category in a fixed amount of time. For example: “name as many animals as possible”. Answers might be “horse, elephant, hippo, zebra, dog and snail”. In general, this task is performed for multiple categories and a composite score is derived. The PVF test is more general compared to the CVF test. An example question for this test might be: “name as many words as possible starting with the letter C” and again answers might be “chocolate, cheque, chloride, cream and car”. Again, the task is performed for multiple letters and a composite score is derived. The PVF test is generally given in either one of two forms: the “CFL” and “FAS” variation with the latter being the most widely used. It is however worth noting that (Barry, Bates and Labouvie, 2008) have observed that participants produce smaller numbers of words on the “CFL” test versus the “FAS”, implying that there is a correlation between difficulty and letter choice.

Matrix reasoning relies on non-verbal reasoning. Participants are given a “matrix” of pictures (i.e. a 2 by 2 or 3 by 3 matrices) that follow a certain visual rule or sequence. One cell in the matrix will be missing and the participants are asked to fill in the missing cell with one of the given possible solutions. By doing so a participant will have to reason “why” the picture fits best into the missing cell. Picture sequences in the matrix may be based on shape and/or colour (Figure 1.4A) but could be extended by including sequences of countable symbols such as dots adding a numeric component (Figure 1.4B).

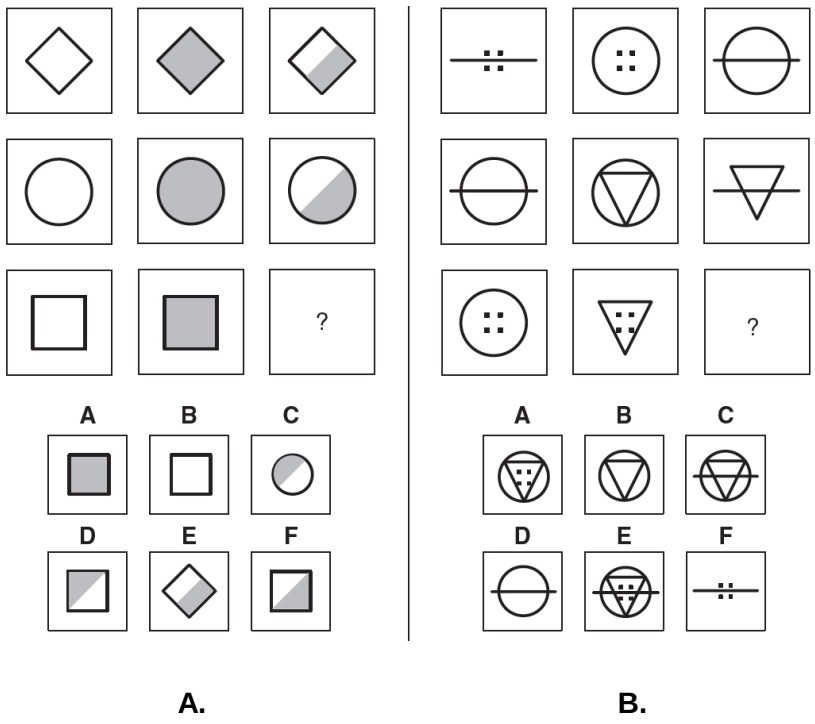


Figure 1.4: Two matrix reasoning test items

Find the correct solution to solve the sequence. Answer 1.4A) = F (all rows same shape, all columns same colour) answer 1.4B) = B (3 sets of 2 pairs per column and per row). Image from Ritchie, 2015.

Processing Speed/attention can be measured in many different ways, but is in general measured by either a "digit-to-symbol" test or by means of a reaction time test. Figure 1.5 shows an example of a "digit-to-symbol" test. For this test participants are given a form containing a template showing digits each with a corresponding symbol (Figure 1.5, above line). Participants are asked to match symbols to digits (Figure 1.5, under line) as quickly as they can in a fixed amount of time.

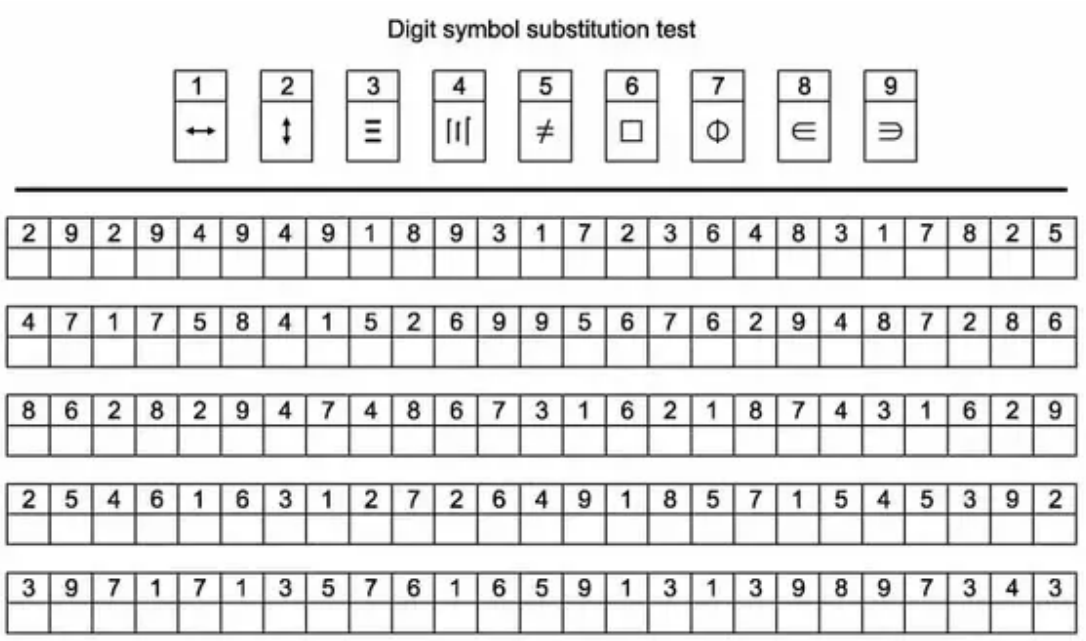


Figure 1.5: Example of digit symbol substitution task.

Image from (Patel and Kurdi, 2015).

The reaction time test measures how quickly a participant can respond to certain stimuli. This can be tested on a simple way by, for example, tapping on a button every time a light comes on. This can be extended by using a touch screen or big board with multiple dots that turn on at random. The idea is to rapidly touch each dot that is illuminated, meaning that participants have to rapidly move their finger/hand to any of the locations.

Language is, as mentioned before, measured using vocabulary. The Mill Hill Vocabulary Scale (MHVS) is a synonym test where participants are asked to find correct synonyms to a given word and to define other words (Raven, Raven and Court, 1988). The words range in difficulty; whereas the first word is common "tomato is a fruit" but thereafter increases in difficulty to i.e. "schooner is a ship".

The reasoning behind this test is that participants will only know the answer if they ever used or seen the words before, making on the spot learning virtually impossible. Another commonly used language test is the National Adult Reading Test (NART; Nelson, 1982). Unlike the written Mill Hill test, the NART is a verbal test were participants are asked to pronounce a list of words that are not pronounced as they are written. Examples of these are: "nausea", "gaoled", "heir" and "hiatus". It is worth noting that a vocabulary test such as the NART depends on the complexity of a language. NART equivalent tests will not work in languages containing only phonologically regular words (i.e., a one-to-one correspondence between writing and pronouncing a word) such is the case in the Spanish language (Cosentino, Manly and Mungas, 2007).

Memory: the digit-span test is one of the many memory tests frequently used in cognitive batteries (Dempster, 1978). Here an examiner will read a list of numbers out loud, and then the participant is asked to repeat this sequence, but in reverse order. The sequences start out short but are increased during each iteration putting more strain on the participants' short-term memory. Another short-term memory test is the *n*-back test (Kane *et al.*, 2007), in this test again a sequence of numbers is read out loud i.e. 2-5-3-9-1-1-7. In contrast to the digit-span test, the participant is asked to only say the number *n* places back from the end of the sequence, i.e. '*n*=4' therefore the correct response is '9'. The commonest memory tests is the word list. Participants asked to remember a list of words and retell them after a certain time either immediately or after a delayed period. The final test that will be discussed is the Wechsler Logical Memory Test. The logical memory test is an extended version of the word list test. The examiner will read a story out loud to the participant which contains a defined number of keywords. Usually this story is two paragraphs long. After the story, the participant is asked to retell the story to the examiner who will count the number of mentioned keywords. These tests can be done immediately after telling the story to test short-term memory but also after a certain amount of time to test long-term memory.

Visual-Spatial Perception: The block design test from the Wechsler Adult Intelligence Scale (WAIS) assesses constructional ability (the ability to perceive, copy or draw shapes, figures, or lines). The block design test requires participants to plan and develop strategies for constructing designs of given blocks. Participants are given a maximum of two minutes to complete each design where each design gets increasingly more difficult to create (Wechsler D., 1998).

The mental rotation task asses the ability to rotate 2D or 3D object mentally. An example can be to give participants a 2D or 3D images of a structure, and the participant is asked to subsequently find the correct structure back while some choices are all different shaped and all differently positioned.

1.3 Literature review of cognitive ability in Major Depressive Disorder

1.3.1 MDD versus controls meta-analysis based studies

The largest-to-date meta-analytic systematic review of cognitive performance in Major Depressive Disorder focussed mainly on seven cognitive performance tests measuring the executive functioning domain but also two non-executive functioning tests (Snyder, 2013). The seven executive functioning tests were inhibition, shifting, updating, verbal and visuospatial working memory, planning and verbal fluency. The two non-executive functioning tests measured vocabulary (language domain) and digit symbol substitution (processing speed, but is also considered to be a component of executive functioning). Snyder. (2013) used 113 studies in their meta-analysis which included in total 7,707 individuals (3,936 MDD patients and 3,771 control participants). Both groups were similar in age and sex: the MDD group contained 2,404 females (61%) with a mean age of 46 years and the control group contained 2,246 females (60%) with a mean age 45 years. Snyder. (2013) observed that MDD patients showed a decrement in phonemic verbal fluency (n=2,850, z=7.72, p<0.01). MDD patients produced significantly fewer words than the control group. Furthermore, MDD patients recoded significantly fewer symbols to digits in the

processing speed test digit symbol substitution, (n=1,904, z=5.25, p<0.01). Vocabulary performance was observed to be lower in MDD patients; however, the effect was not significant (n=2,175, z=2.68, p=0.07).

Lim *et al.* (2013) conducted the largest meta-analysis study of logical memory (measuring verbal declarative memory) to date (n logical memory immediate=291; n logical memory delayed=348). They observed that MDD patients performed significantly less well than matched controls on both logical memory immediate and delayed. This result has been previously reported by smaller studies such as Delgado, Kapczinski and Chaves (2012) (n=29) and Maeshima *et al.* (2013) (n=67) both not included in the Lim *et al.* (2013) meta-analysis, who used studies published before 2011. Travis *et al.* (2014), also not included in Lim *et al.* (2013), observed no significant difference in logical memory immediate and delayed performance between MDD patients and controls, however the sample size was very small (n=15). Significant decrements were also observed in the attention domain, via the digit span test and continuous performance test where MDD patients performed slower compared to controls. The final domain examined, visuospatial processing (immediate and delayed visual memory), did not report differences between MDD patients and controls (Lim *et al.*, 2013).

1.3.2 MDD versus controls non-meta-analysis based studies

Cullen *et al.* (2015) is the largest non-meta-analysis study to date investigating cognitive performance between adults with and without a lifetime history of mood disorder features. Adults were assessed aged between 40 and 69 years, registered with the National Health Service (NHS) and lived within 25 miles of a UK Biobank study assessment centre. 172,745 participants filled in the mood disorder questions. After excluding individuals with missing data or self-reported neurological disorders that can impair cognitive performance (e.g. Parkinson's disease) a study population of 143,828 individuals was left. Of these, 111,960 individuals formed the control group (note that 86,190 individuals had no clinically significant mood disorder

features, the remaining 23,384 individuals met a small number of criteria which was not enough for a depression classification and 2,386 had manic symptoms that did not fulfil the criteria); 7,607 individuals met the criteria for single-episode; 14,386 individuals for moderate-recurrent and 8,354 individuals for severe-recurrent major depressive disorder. The remaining 1,521 individuals fell in the bipolar disorder group. The study tested five cognitive domains. The assessment included: reasoning (verbal and numeric reasoning), reaction time (press button when observing a matching pair of symbols), numeric memory (immediate reverse recall of a string of numbers), pairs matching (pairs matching test) and prospective memory (delayed recall of given task). Linear models were performed between cognitive performance of named tests and disease status adjusted for age, sex, smoking (current versus ex or never smoked), alcohol (daily/almost daily versus ex or never drank alcohol), on psychotropic medication, current depressive symptoms score, education (university degree – yes/no), and Townsend score (socioeconomic status). Cullen *et al.* (2015) observed that all three major depression groups (single episode, recurrent moderate and recurrent severe MDD, in that order) significantly outperformed the control group in the tests measuring prospective memory (*OR*=1.37 and 1.25, p<0.001; and 1.11, *p*=0.001) and reasoning (*OR*=0.26, 0.22 and 0.12, *p*<0.001). Reaction time was significantly shorter in single episode (*β*=-5.47 ms, *p*<0.001) and moderate recurrent depression (*β*=-6.24, p<0.001) groups compared to controls. However, there was no difference observed between severe recurrent depression (*β*=-0.82, *p*<0.530) and controls. Numerical memory performance was significantly better in single episode (*β*=0.07, *p*=0.004) and moderate recurrent depression groups (*β*=0.06, *p*=0.004). Finally, fewer errors were made during the pairs matching test in the single episode (*β*=0.98, *p*=0.009) and severe recurrent depression group (*β*=1.03, *p*<0.001); surprisingly this was not observed in the moderate recurrent group (*β*=0.99, *p*=0.428) which falls between previous mentioned groups.

Halvorsen *et al.* (2012) is one of the few studies that examined cognitive performance between healthy controls (*n*=50) and currently depressed (*n*=37) and previously depressed (*n*=81) individuals. This has not been done previously and is therefore

interesting but small of size. Halvorsen *et al.* (2012) compared the performance of 19 neuropsychological tests including digit symbol coding and verbal fluency. Missing test data was replaced by group mean scores and all data were analysed using MANOVA. The main findings indicated that cognitive performance does not significantly differ between healthy controls and currently depressed or healthy controls and previously depressed individuals. A mild and limited decrement was observed in processing speed (digit span backwards test) and working memory in the currently depressed group however this was not significant after using a Bonferroni-adjusted alpha level of 0.002.

1.3.3 Single versus recurrent (or mild versus severe)

Differences in cognitive performance between single-episode and recurrent MDD have not been studied as widely as cognitive performance differences between MDD patients and healthy controls. Talarowska, Zajaczkowska and Galecki, (2015) studied the cognitive performance of 210 patients with MDD (*n* single-episode=60, *n* recurrent=150) and observed that the cognitive domains of executive functioning, memory and processing speed showed significant decrements in recurrent MDD in relation to single MDD. Cognitive differences between single and recurrent depression was not studied directly in the UK Biobank study (Cullen *et al.*, 2015). Halvorsen *et al.* (2012) also investigated whether individuals differed in cognitive performance in a currently (*n*=37) versus previously depressed (*n*=81) study design. Severity of MDD was negatively correlated with processing speed tests such as digit span forward (r=–0.20; p=0.029) and backward (r=–0.20; p=0.028).

1.4 Overview of genetic epidemiology

The field of genetic epidemiology is a specific field within epidemiology. Epidemiology is in general defined as “the study of the distribution, determinants of health-related states and events in populations” (Porta and International Epidemiological Association., 2008). Genetic epidemiology investigates the role of genetic factors and helps determining diseases and health in families and in population.

Note this section only discusses study designs used in the field of genetic epidemiology. Study designs specific to MDD and cognitive performance are discussed in section 1.6 and 1.7.

1.4.1 Twin studies and familial aggregation

Twin studies are at the basis of genetic epidemiology and reveal the importance of environmental and genetic influences on traits (Plomin, DeFries and McClearn, 1990). There are two types of twins: monozygotic twins, also known as identical twins, and dizygotic twins, also known as fraternal twins. Monozygotic twins develop from a single egg fertilised by a single sperm which splits into two. Because the egg splits after fertilisation monozygotic twins share 100% of their DNA and are therefore genetically identical. Dizygotic twins however develop from two separate eggs fertilised by two separate sperm therefore dizygotic twins share around 50% of DNA on average which is equal as with any other sibling pair.

Monozygotic and dizygotic twins typically share the same postnatal environment as they are raised by the same parents at the same time and i.e. eat the same food and play with the same toys.

Often the assumption is made that due to sharing the same womb at the same time twins also shared the same prenatal environment i.e. stress and/or smoking of the mother. However, this is not always the case and depends, in part, on whether twins have shared the same most outer membrane called the chorionic sac (chorionicity). As dizygotic twins come from two different fertilised eggs they will not share a chorion and develop individual placentas. Monozygotic twins can be monochorionic and share a placenta or dichorionic and have separate placentas like dizygotic twins. Studies have shown weight differences between monochorionic and dichorionic monozygotic twins, implying that prenatal environment should not be assumed to be the same (Marceau *et al.,* 2016). Regular siblings share the same amount of DNA on

average as dizygotic twins but do not necessarily share the same pre- and post-natal environment, i.e. different approach of raising a child, moved to a different house or different food choice.

Most phenotypes can be modelled as the sum of genetic (nature) and environmental (nurture) factors. Twin studies are useful to study the contribution of nature and nurture and the affect one has on a phenotype without the other. In a twin study, the concordance of a phenotype can be compared between monozygotic and dizygotic twins. This allows one to estimate the effect of genes while the environment is held constant. If a phenotype is primarily genetic, a study will show that a phenotype has a higher concordance in monozygotic twins than dizygotic twins, as the latter shares less DNA. However, if a phenotype is mostly environmentally-driven, a study would show a similar concordance of a phenotype between both twin groups (Plomin, DeFries and McClearn, 1990). Critics claims that because twins are not a random sample of the population they do not represent the general population. The most basic problem in twin studies is that resemblance of a trait is due to having a shared environment therefore inflating the estimated heritability compared to the actual heritability. Adoption studies are a good way to assess the effect of non-shared environment. In adoption studies the phenotype of the adoptees is compared with phenotype of the adoptive versus the biological parent (Plomin *et al.*, 1997). Genetically related individuals separated from birth and raised in different and uncorrelated environments will resemble each other due to genetic reasons. Genetically unrelated individuals e.g. adopted siblings will resemble each other due to shared environmental factors (Plomin and Daniels, 2011).

Familial aggregation studies are a common method of choice in genetic epidemiology studies to determine a possible genetic aetiology of phenotype. The rationale behind familial aggregation is to identify a proband (the individual who was first ascertained into the study) with a specific phenotype and determine whether relatives have an excess frequency of the same phenotype (Matthews, Finkelstein

and Betensky, 2008). Risk of a phenotype within a family is often calculated by means of a simple division of the amount of individuals having the phenotype by the amount of individuals not having the phenotype all members within the same family. This represents the probability of having a phenotype within a family. An odds of 1 indicated that the family contains as many individuals with and without the phenotype. An odds <1 indicates that more individuals in the family do not have the phenotype of interests while >1 indicates that more individuals in the family have the phenotype. A phenotype has an excessive frequency in a family when the $odds_{family}$ is larger than an odds of appropriate reference population (null population). Because related individuals have more DNA in common and share more the same environment than randomly selected reference individuals the conclusion can be drawn that the excess phenotype has a genetic basis.

1.4.2 Heritability

Heritability or h^2 is the proportion of variance in a phenotype that is attributable to genetic differences. A common misinterpretation is that heritability is the percentage of genetic factors making up a phenotype. There are multiple ways to calculate heritably depending on the data at hand and relatedness of individuals.

As mentioned before environmental factors are kept constant in both mono and dizygotic twins; therefore, twins are ideal for estimating heritability. Let us assume we observe a correlation (*r*) between schizophrenia in monozygotic twins of 0.7 but we observe a correlation between dizygotic twins of 0.4. This can be added to '*Falconer's Formula*' to assess the genetic heritability of schizophrenia (Falconer and Mackay, 1996) (Equation 1.1).

$$h^2 = 2\left(r_{monozygotic} - r_{dizygotic}\right) \quad (1.1)$$

Using the numbers above 0.6 or 60% of variation in schizophrenia is due to having different genes or alleles. Note that the difference in correlation between mono and dizygotic twins is multiplied by two due to monozygotic twins sharing twice more DNA than dizygotic twins.

Phenotypic variance (V_p) can be broken down into the sum of multiple components: genetic variance (V_G), environmental variance (V_E) and genetic-environmental Interaction (V_{GE}). Where genetic variance can be broken down to the sum of: genetic additive variance (V_A), dominance genetic variance (V_D) and genetic interaction variance (V_I).

Broad sense heritability (Equation 1.2) reflects all the genetic contributions (V_G=V_A+V_D+V_I) to a population's phenotypic variance (Kempthrone, 1957).

$$H^2 = \frac{V_G}{V_p} \quad (1.2)$$

While narrow sense heritability (Equation 1.3) shows the amount of phenotypic variation that can be contributed due to only additive variance (V_A).

$$h^2 = \frac{V_A}{V_p} \quad (1.3)$$

Heritability does not indicate the degree to which the phenotype is genetically determined. A heritability score indicates the genetics variance in a phenotype in a population and cannot be translated to individual characteristics. Also, every

population is different therefore every no universal heritability exists. Even when heritability is high, environmental factors may influence a phenotype.

1.4.3 Genetic architecture

Genetic architecture refers to the underlying genetic template of a (human) trait and the properties relating to the handing of variation to the next generation within a population. The detection of the genetic architecture depends in part on both the penetrance (proportion of individuals carrying a certain variant associated with the trait) and effect size (magnitude of the allelic effect on the trait). Identifying high penetrance variants works well in family based studies where most related individuals share the same trait/disorder. Family based studies are almost incapable of detecting low penetrance variants in sporadic traits and disorders as it becomes unclear whether a variant is causal or simply due to relatedness. For these low penetrance traits/disorders a large sample size is required of non-related individuals.

1.4.4 Linkage studies

According to the Law of Independent Assortment (Mendel's second law) alleles for separate traits are passed independently of one another, however Genetic linkage violates this law. Linkage studies are used to associate phenotypes to genomic variants. If a phenotype is common in a family (high penetrance) along with specific genomic markers (variation only or often found in individuals with a specific phenotype) the conclusion can be drawn that the variation responsible for the phenotype are either the markers or markers located physically close to these markers on the chromosome (Pulst, 1999). Linkage studies can be split into parametric and non-parametric linkage. Parametric linkage is the most commonly used method. During parametric linkage the probability of a gene being important/associated with the phenotype of interest is calculated by the LOD ($\log_{10}$ of the odds ratio scores). Using a pedigree, the LOD indicates the probability where the phenotype and a genetic marker are inherited together due to linkage. Non-parametric linkage analysis, in turn, studies the probability of an allele being identical

by descent (IBD) (Nyholt, 2000). Two alleles at a locus are IBD if and only if the two alleles are both descendants of a common ancestral allele. Non parametric methods test for more sharing in groups with a phenotype than one would expect when there is no linkage. Due to this structure non-parametric methods use small, nuclear families in contrast to the large multigenerational families parametric linkage uses (Kruglyak *et al.*, 1996).

1.4.5 Candidate gene association studies

This section contains study design examples online. For many years candidate gene studies were at the forefront of genetic association studies, i.e. identifying risk variants associated with a particular disease. Candidate gene studies focus on the selection of specific genes with prior knowledge about gene function and their relationship with the phenotype of interest. Candidate gene studies generally follow 3 steps:

1. Selecting a putative candidate gene based on its relevance in the mechanism of the phenotype of interest.

2. Assigning and selecting SNPs, usually physically located in and up and downstream from the gene.

3. The gene variants are tested for association with the phenotype by observing its occurrence in cases with the disorder or who have the phenotype and control subjects which do not. This step is not limited to binary outcomes and can also be applied to quantitative outcomes.

This approach increases the knowledge of specific genes and may be clinically relevant as a potential disease diagnostic tool. However candidate studies have also been criticised for their low rate of replication, incorporation of a priori knowledge and lack of causality (Tabor, Risch and Myers, 2002).

1.4.6 Genome-Wide Association Study and Polygenic Scores

1.4.6.1 Genome-wide association study

A genome-wide association study (GWAS) is a widely used approach for examination of common genetic variants called Single Nucleotide Polymorphisms (SNPs) with a phenotype or disorder (Tabor, Risch and Myers, 2002). The idea behind GWAS is to test thousands to millions of SNPs spread over the genome to detect genetic association with a phenotype. However, one needs to take Linkage disequilibrium (LD) into account. LD is the non-random association of alleles at different loci in a given population. A genotyped SNP that is in LD with an unknown causal variant will follow roughly the same allelic distribution and will most likely show the same association score during a GWAS. Therefore, even if the causal variant of the phenotype is not present in the dataset, due to linkage disequilibrium it is expected that the association signal will spread through its surrounding genomic region. The statistical association between genotypes and phenotypes can be done by means of a χ^2 test, Fisher exact test or logistic regression analysis when the outcome is binary (1 = 'present' and 0 = 'not present') or by means of a t-test, linear regression analysis, ANOVA or maximum likelihood when the outcome is quantitative (Table 1.3).

Individual	Phenotype (binary)	Phenotype (quantitative)	Sex	SNP 1	SNP 2	…	SNP n
1	1	0.5	F	2	1	…	0
2	1	-1.3	M	1	2	…	1
3	0	2.99	F	0	2	…	1
…	…	…	…	…	…	…	…
m	0	-10.36	M	1	-9	…	2

Table 1.3: Small representation of GWAS data.

Column 1 represents the individual ID, 2 the phenotype if binary, 3 the phenotype if quantitative, column 4 till n represents the genotype for every measured SNP (0 = homozygote allele one, 1 = heterozygote, 2 = homozygote allele two and -9 is missing data).

Finally, the p-values derived from either a logistic or linear regression analysis explaining the probability of observing the result between a SNP and the phenotype assuming no association. This can be presented in a Manhattan plot (Gibson, 2010;

Witte, 2010) in order to highlight the genomic regions harbouring an excess of statistically significant SNPs (Figure 1.6).

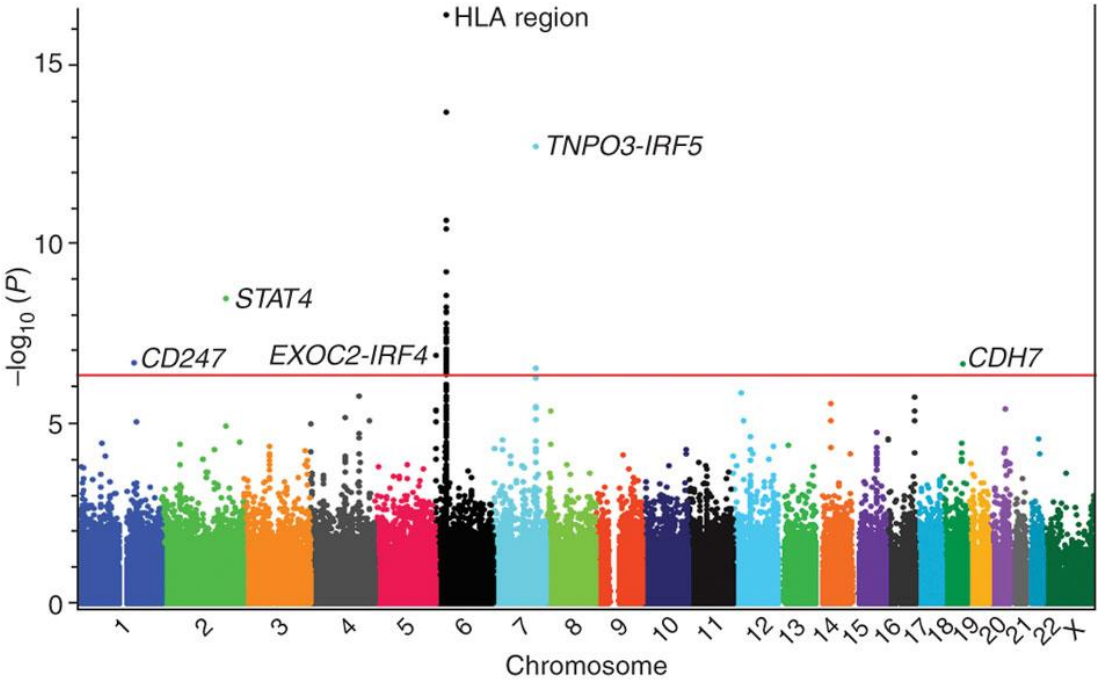


Figure 1.6: Example of a Manhattan plot.

On the y-axis the statistical association between variations present at each SNP (shown as a dot) and systemic sclerosis. X-axis depicts the genomic position of each SNP (chromosome and physical position on the chromosome). The red line indicates the (putative) threshold for statistical association. Image from Radstake et al. (2010).

Statistical significance is determined by means of a conservative Bonferroni threshold to account for multiple testing. This threshold is derived by taking an arbitrary *p*-value threshold, e.g. 0.05, and dividing this number by the number of independent tests (SNPs) performed. In most GWA studies this threshold lies in the $5x10^{-08}$ region.

1.4.6.2 Polygenic Score

GWAS aims to detect associations between single SNPs and the outcome, making it a single-locus method. The polygenic score is an additive method meaning that it

looks at the combined effect of a range of SNPs and their association with the phenotype. Polygenic scores utilise summary output from genome-wide association studies to create scores in independent target datasets. Polygenic scores are calculated as following: first, the SNPs in the independent target dataset GWAS are ranked based on the observed p-values (Figure 1.7A). Next, for each SNP the number of copies of the risk allele for each individual in the discovery dataset is determined (Figure 1.7B) and multiplied by the regression weight (i.e. beta coefficient or odds ratio) from the initial GWAS (Figure 1.7C). The scores are summed over all SNPs in each individual falling under an arbitrary p-value threshold to give a single number. This applies solely when no genotype has any missing value in any individual, e.g., when using imputed data (the statistical inference of unobserved genotypes). For genotype data with missing values, the polygenic score is divided by the number of SNPs used to calculate the polygenic score to get a weighted average.

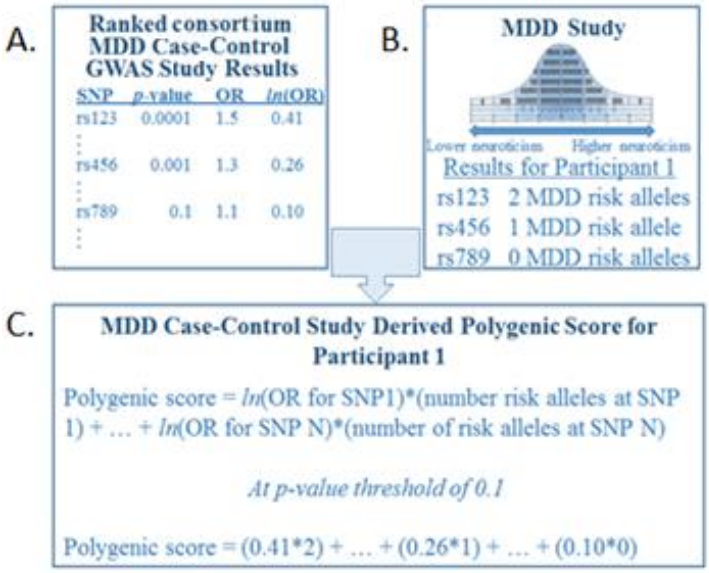


Figure 1.7: Step by step approach for the calculation of polygenic scores.

Using summary statistics from GWAS data (A) and genotype data (B). Image taken from Rosetrees Trust Biomedical Research Grant 2014, Understanding the Genomics of Cognitive Deficits in Major Depressive Disorder through Novel Statistical Models, K. Nicodemus.

This score can be used as a predictor of a trait of interest. The amount of variation explained by the polygenic score can be calculated by means of R^2 (continuous outcomes) or Nagelkerke's R (binary outcomes).

1.4.6.3 Review of epistasis

The definition of the term epistasis has changed many times but is commonly referred to as "gene-gene interaction". Historically, epistasis was defined as the masking or modifying effect one allele has over another allele at a different locus (Bateson and Mendel, 1909). This was later extended more quantitatively by Fisher (1919) as "deviation from additivity of two genetic variants on a phenotypic trait". We know that genes are the blueprint of proteins and proteins act together in networks and multiple networks enact biological functions. Because of this, most epistatic studies have focussed their attention on within-network dependencies in non-human model organisms (Brockmann *et al.*, 2000; Cheng *et al.*, 2011; Huang *et al.*, 2012; Mackay, 2014; Grice, Liu and Webber, 2015; He *et al.*, 2016). However, due to the requirement of multiple proteins, the consequence of one genetic variant affecting the functioning of one protein in a network may not be independent of variants affecting other network members; in other words, networks often have redundancies that buffer the network when small changes occur. When this leads to a measurable outcome these dependencies may condition the outcome effect of a specific variant. This form of epistasis can be split into antagonistic (where a combination of alleles together diminishes the effect of each allele individually) or synergistic (where a joint effect of the alleles exacerbates the effect of each allele individually) epistasis.

Epistasis has been observed and documented in multiple non-human organisms i.e. drosophila (Huang *et al.*, 2012; Grice, Liu and Webber, 2015; He *et al.*, 2016), mice (Brockmann *et al.*, 2000; Cheng *et al.*, 2011; Mackay, 2014) and other model organisms (Mackay, 2014). Whether epistasis is an important feature of the architecture of human traits or disorders is currently unclear.

Hill, Goddard and Visscher (2008) performed an extensive evaluation of evidence from empirical studies of genetic variance components (additive, dominance and epistatic). The study argues that epistasis does not occur in humans and previously observed results are the result of additivity of genetic effects. The evidence of this argument was observed in comparing the phenotypic correlations between monozygotic (MZ) and dizygotic (DZ) twins. As expected, on average the correlation between MZ twins is about twice as high as DZ twins in a wide range of phenotypes. According to (Hill, Goddard and Visscher, 2008) this can be explained by additive variance as it is highly unlikely that the variance due to common environmental factors, assortative mating and non-additive genetic factors cancel each other out by pure chance. Using additive models Hill, Goddard and Visscher (2008) were able to show that most genetic variance appeared to be additive. Given these results, it is possible that dominance and epistatic interactions are almost non-existent. Another possibility is that these results are mainly driven by the fact that most allele frequencies are distributed towards extreme values ("U" shaped distribution) leading either to a high or low additive variability.

Huang and Mackay (2016) are not in agreement with Hill, Goddard and Visscher (2008)'s claim that epistasis is unimportant due to it increasing the additive variation and this is the variation that drives correlations between relatives. Huang and Mackay (2016) applied different parameter settings used for assessing genetic variance that leads to large proportions of genetic variance due to non-additive factors. Furthermore, the implications of the lack of correspondence between homozygous, heterozygous and epistatic interaction effects and additive, dominance and interaction variance components were discussed. The study showed that when applying alternative parameters for assessing variation due to dominance and additive × additive effects, the majority of total genetic variation can be captured. Therefore, one could say that dominance and additive variation using standard settings are non-important. The study showed that neither the standard nor alternative parameters of assessing genetic variance shows any information regarding whether or not the total amount of genetic variation leans towards either dominance, additive or epistasis. In short, when using an additive model evidence for solely additive variance will be found; however, the

use of dominance or epistatic models will provide evidence for those effects. Many statistical genetic approaches downplay the contribution epistasis has on the total amount of genetic variation by testing for epistasis after testing for additive and dominance components first therefore inflating their contribution (Sackton and Hartl, 2016; Webber, 2017). As the true effects underlying the genomic architecture of human complex traits is unknown, models need to be developed that can test for additive and epistatic effects simultaneously, thus providing the ability to evaluate the relative importance of each.

1.4.6.4 Review of Nicodemus et al., 2014

Nicodemus *et al.* (2014) introduced a novel statistical model that incorporates single gene, polygenic and epistatic components to assess the association between SNPs in the *ZNF804A* pathway and cognitive performance in psychosis. The study used genes in the *ZNF804A* pathway (*A2M*, *ACTG2*, *C2RF80* amongst others) defined by Hill *et al.* (2012) and subsequently selected all SNPs in a +/- 20kb region of *ZNF804A* pathway genes. Polygenic scores were calculated for three *p*-value ranges $p < 10^{-5}$ (*n* SNPs = 10), $p < 0.05$ (*n* SNPs = 218) and $p < 0.5$ (*n* SNPs = 1525) using the Psychiatric Genomics Consortium 1 (PGC) schizophrenia genome-wide association study.

Derived polygenic scores were regressed against seven cognitive measures: IQ, episodic memory, working and spatial working memory, attention and social cognition in a narrow psychosis (Schizophrenia and Schizoaffective disorder individuals) and broad psychosis (narrow psychosis set + Bipolar disorder + MDD + psychosis not otherwise specified individuals) set. Among both narrow and broad psychosis group a higher *ZNF804A* derived polygenic score (range $p=0.5$) was associated with poorer performance in spatial working memory.

Two-SNP interaction modelling was conducted for all SNPs in the polygenic range ($n=218$) across 100 bootstrap samples of half of the narrow psychosis (training data) to test for epistasis. The median of the *p*-value from the 100 samples was taken. This resulted in a regression model that contained the polygenic score and two two-SNP

interaction terms. When this model was used to predict performance in spatial working memory in the two independent samples of psychosis cases it increased the variation explained (R^2) from 1.3% using only the polygenic score to between 4.8-6.2%, an increase of between 3.5-4.9%.

It is worth noting that this model is still relatively simplistic in modelling the genetic architecture of complex traits; in particular, the epistatic component as the model only allowed for pairwise interactions between SNPs. A more flexible specification of the epistatic component may be given by the use of non-parametric logic, classification and regression trees within a regression model.

1.5 Genetic epidemiology of Major Depressive Disorder

1.5.1 Heritability, familial aggregation and twin studies

Kendler *et al.* (2006a) studied the heritability of MDD in 42,161 Swedish twins which included 15,493 complete pairs and 11,175 twins whose co-twins was not assessed. The 15,493 complete twin pairs were divided in five groups based on sex and twin type (Figure 1.8).

Sex and Zygosity	Number of Complete Twin Pairs	Correlation Between Twins for Liability to Lifetime Major Depression	
		Tetrachoric Correlation	95% CI
Female-female, monozygotic	2,317	0.44	0.38–0.50
Female-female, dizygotic	3,185	0.16	0.10–0.22
Male-male, monozygotic	1,774	0.31	0.20–0.41
Male-male, dizygotic	2,584	0.11	0.01–0.20
Male-female, dizygotic	5,633	0.11	0.05–0.16

Figure 1.8: Sample size twin pairs and correlation between twins and liability for lifetime depression. *Images from* (Kendler *et al.*, 2006a).

Tetrachoric correlations (correlation between two theorised normally distributed continuous latent variables, from observed ordinal variables) and 95% confidence

intervals (CIs) are seen for the five twin zygosity groups in Figure 1.8. The correlations within MDD monozygotic twins are substantially higher than within dizygotic pairs. Next, the correlations in the same-sex female pairs exceed those seen in the same-sex male pairs. Finally, MDD in non-same sex dizygotic pairs is the same as the same-sex male dizygotic pairs. Kendler *et al.* (2006a) found no evidence for shared environmental risk factors and their importance in relation to MDD between sexes but observed that the heritability of liability of MDD was greater in women than men. The genetic correlation (the proportion of variance that two traits or groups share due to genetic causes) of liability of MDD between men and women was estimated at 0.63, implying a substantial proportion of sex specific MDD genetic risk factors. The twin based heritability of MDD was estimated to be 0.38 or 38%.

Sullivan, Neale and Kendler (2000) selected five family studies from two countries for familial aggregation analysis. In this study, probands (a person serving as the starting point, often with the phenotype of interest) with MDD were matched with controls without MDD based on sex and age to estimate the prevalence of MDD in related individuals (first degree relatives, either a parent, full sibling, or child). Sullivan, Neale and Kendler (2000) showed that there was strong evidence of an association between MDD in probands and MDD in first degree relatives. The odds ratios (OR=2.84, 95% CI=2.31–3.49) were homogeneous across all five studies, therefore in aggregate these five studies provide consistent evidence in support of the familiality of MDD. It has to be noted that some studies were sampled from clinical populations (Gershon, 1982) in contrast to general populations therefore the overall odds ratio may be biased. Due to this suggested bias, heritability of MDD in clinical and general studies were calculated separately. Estimates of heritability ranged between 37% in general population studies to 43% in clinical studies. Fernandez-Pujals *et al.* (2015) derived heritability estimates from a large Scottish based population study and observed a heritability of 28%. This estimate is lower than Kendler *et al.* (2006) who estimated a 37% twin heritability. This was to be expected as twin heritability estimates are in general higher than pedigree based estimates (Pilia *et al.*, 2006). Large family studies allow for more power to detect shared

familial environment effects compared to twin studies due to the fact that in twin studies only two individuals per family are generally observed. Family environment effects are present in twins but are not statistically different from zero and are therefore dropped from the model, thus heritability may be upwardly biased.

Lubke *et al.* (2012) estimated heritability of MDD due to genotypic variation (SNPs) between Dutch twins. Lubke *et al.* (2012) compared two methods(Yang *et al.*, 2010; So, Li and Sham, 2011). Yang *et al.*, 2010) calculated the variance of MDD by decomposing the trait into an additive effect of all SNPs and a residual component that is due to random noise/unmeasured environmental influences and/or unmeasured genetic variants. The method proposed by So, Li and Sham (2011) is entirely different, and is applied subsequent to a GWAS. The overarching idea is to compare the distribution of *z*-statistics of the regression coefficients of genome-wide SNPs in a GWAS to the theoretical null distribution of *z*-statistics representing no effects. The additive genetic variation of MDD due to all SNPs was estimated at 32% (Yang *et al.*, 2010) and 28% (So, Li and Sham, 2011). This is much larger than the 1% of variation explained by GWAS (Demirkan *et al.*, 2011). The difference can be explained by multiple factors. First, the effect size of involved SNPs might be small, resulting in insufficient detection power for a GWAS and secondly, methods used by (Lubke *et al.*, 2011) include data-mining procedures that can efficiently extract information that is present in the genome wide SNP data.

1.5.2 Linkage studies

Flint and Kendler (2014) reviewed eleven MDD linkage studies and reported the commonly used logarithm of odds ratio (LOD) score. Most studies included in the review used MDD affected siblings. In the study, a LOD score of 2.2 was assumed to suggest evidence of linkage, a LOD score surpassing 3.6 was associated with significant linkage and 5.4 with highly significant linkage. Flint and Kendler (2014) concluded that heterogeneity was clear between studies, with one study reporting more loci at higher levels of significance than others (Zubenko *et al.*, 2003). It is

worth noting that Zubenko *et al.* (2003) reported unusually low simulation-based LOD score thresholds for analyses without covariates. In addition, multiple publications reported overlapping datasets leading to inflation.

1.5.3 Replicated candidate genes for MDD

Bosker *et al.* (2011) performed an extensive analysis trying to replicate candidate genes for major depressive disorder using genome-wide association data. Bosker *et al.* (2011) observed 78 papers resulting in 57 genes reported to be statistically significant different between MDD cases and healthy controls e.g. *DISC1* and *COMT*. Ninety-three SNPs were mapped to the candidate genes and tested for replication in the Genetic Association Information Network (GAIN) MDD dataset (1,738 cases and 1,802 controls) (http://www.fnih.org/GAIN). Four out of the fifty-five candidate genes were found to have an association with MDD during replication (before correcting for multiple testing): *C5orf20* (rs12520799; p=0.038), *NPY* (rs16139; p=0.034), *TNF* (rs76917, p=0.0034) and *SLC6A2* (multiple SNPS, p=0.039) with *TNF* (*rs76917*, p=0.0034) being identified as the only gene remaining genome-wide significant. Luo *et al.* (2016) conducted a systematic literature search to find genetic case-control association studies on MDD, published between September 1st, 2007, the end search date in Bosker *et al.* (2011), and June 10th, 2012. This resulted in 157 articles investigating candidate gene associations with MDD. Of these 157 articles, 81 reported significant associations (p<0.05) resulting in 201 SNPs observed in 97 candidate genes. Of these 185 SNPs in 89 genes could be mapped and were tested for replication using data from the GAIN genome-wide association study (MDD: n=1,352; chronic MDD subsample: n=225; controls: n=1649). Nine candidate SNPs in eight genes for MDD were replicated (*PSMB4, ADK, POMC, HTR1A, PCLO, CDC42SE2, SIRT1, and SLC29A3*). Six SNPs in five genes were significantly associated with severe and chronic MDD (*PSMB4, ADK, POMC, HTR1A, and PDE4B*). 18 genes were significantly associated with MDD on a gene level. These genes either contained significantly higher numbers of significantly associated SNPs or insertions/deletions than expected by chance (n=13) or contained SNPs that were significantly different from the total amount of SNPs (n=7). No

candidate genes were replicated. None of these candidate genes are observed in the largest MDD GWAS (Wray and Sullivan, 2018).

1.5.4 Large GWAS consortia

The Psychiatric Genomics Consortium (PGC) MDD Working Group is a multi-national study effort to conduct meta- and mega analyses (joint analysis of participant data from multiple available studies) of genome wide data for psychiatric disorders. Ripke *et al.* (2013) published a mega analysis of GWAS studies of MDD also commonly referred to as PGC-MDD 1. Genotypic and phenotypic data were provided by multiple groups and uniform quality control was performed. The total dataset consisted of 18,759 unrelated and independent individuals all from a European ancestral background (9,240 MDD cases and 9,519 controls). Logistic regression was performed to test for associations between MDD and 1,235,109 imputed autosomal SNPs controlling for 51 covariates. Even with this large number of cases and controls no genome-wide significant association was observed between MDD and any SNP (Figure 1.9).

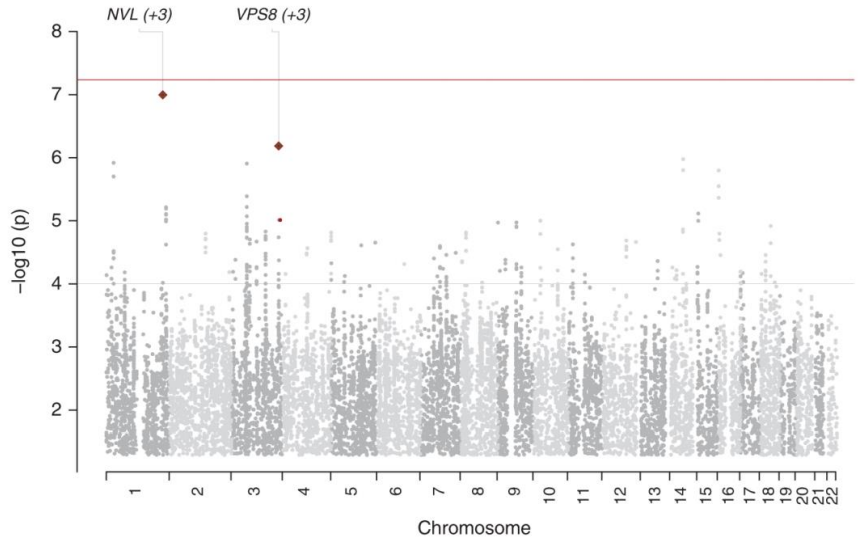


Figure 1.9: GWAS results PGC1 MDD analysis

No SNP surpassed the significance threshold (red line). 2 SNPs with the strongest non-significant association are highlighted. Images taken from (Ripke et al., 2013).

China, Oxford and Virginia Commonwealth University Experimental Research on Genetic Epidemiology (CONVERGE) is another large consortium investigating the genetic underpinnings of depression which does not limit itself to investigating solely European ancestry individuals. Cai *et al.* (2015) recruited 11,670 Han Chinese women through a collaboration involving 58 hospitals in China. After initial quality control 10,640 samples (5,303 cases of MDD, 5,337 controls) and 6,242,619 SNPs for inclusion in genome-wide association studies were used. Two loci exceeded genome-wide significance in association with MDD after controlling for multiple testing: one located in the SIRT1 gene on chromosome 10 (SNP = rs12415800, chromosome 10, $p=1.92 \times 10^{-8}$), and the other in an intron of the LHPP gene (SNP = rs35936514, chromosome 10, $p=1.27 \times 10^{-8}$) (Figure 1.10).

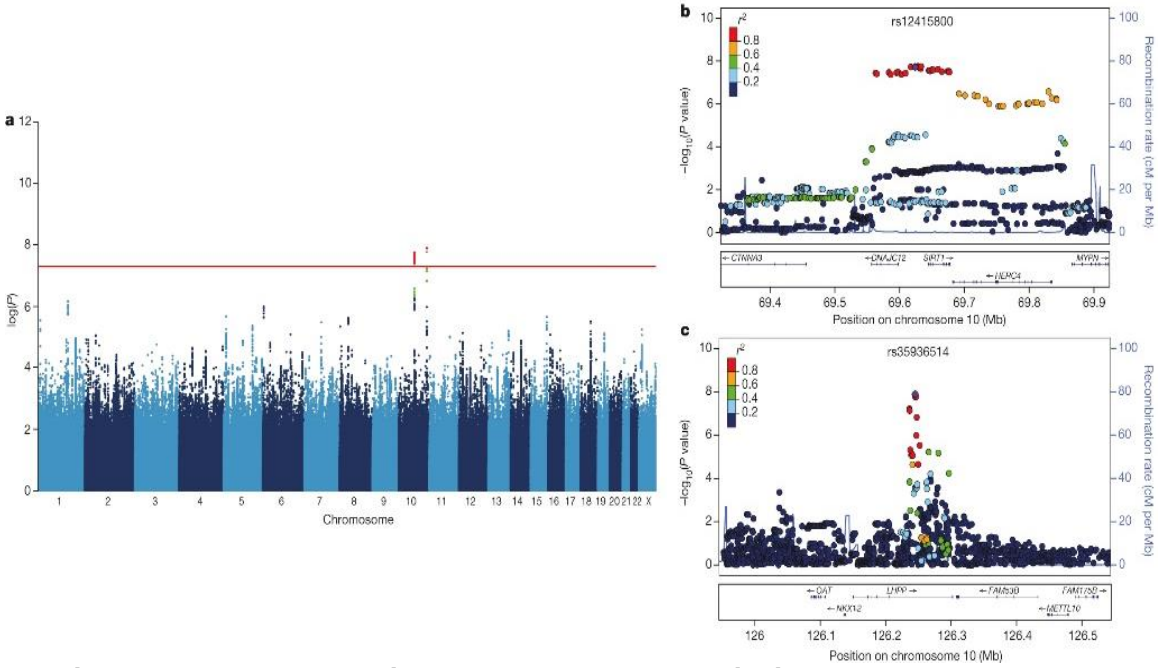


Figure 1.10: GWAS results CONVERGE MDD analysis.

2 SNP surpassed the significance threshold (red line). Two SNPs with the strongest non-significant association are highlighted. Both SNPs are shown on plots right side. LD structure is shown between both SNPs and their surrounding SNPs. Red indicates r^2 0.8 while dark blue indicates 0.2. Images taken from Cai et al. (2015).

Howard *et al.* (2017) performed the largest to date analysis of depression using a single population based cohort. The study performed multiple genome wide association studies on 331,374 unrelated individuals of the UK Biobank dataset (*n*=488,380).

Compared to other studies Howard *et al.* (2017) used three definitions of depression i.e. broad (*n*=113,769, prevalence=35.27%), probable (*n*=30,603, prevalence=17.54%) and ICD-code (*n*=8,276, prevalence=3.80%) depression. Where broad depression was defined as having seen a general practitioner for nerves, anxiety, tension or depression. Probable depression was present when participants were either: "Depressed/down for a whole week; plus at least two weeks duration; plus ever seen a general practitioner or psychiatrist for nerves, anxiety, or depression" or "ever anhedonia for a whole week; plus at least two weeks duration; plus ever seen a GP or psychiatrist for nerves, anxiety, or depression". ICD-coded depression was derived from linked hospital admission records. Positive classification for depression was made if participants had either an ICD-10 primary or secondary diagnosis for a mood disorder. GWAS, controlling for sex, age, genotyping array and eight principal components, for all three forms of depression yielded 17 genome-wide significant variants for broad depression, 5 for probable depression and 1 for ICD-10 depression. Of these 4 (rs6699744, rs9530139, rs40465 and rs68141011) were involved with brain expressed genes (*RPL31P12*, *B3GALTL*, *ZNF391*, *ZNF204P*, *ZNF192P1*, *ZSCAN31* and *ZSCAN23*).

1.5.5 Review of neuroticism

Neuroticism is a heritable and moderately stable (stability of individual differences) personality trait characterised as a tendency to respond with a negative emotional response to threat, frustration, or loss (Matthews, Deary and Whiteman, 2009). Neuroticism has shown to be important in public health research due to it being correlated with a wide range of mental and physical traits (Lahey, 2009). Studies have consistently shown a strong positive association between higher neuroticism scores and MDD (Muris *et al.*, 2005; Chan, Goodwin and Harmer, 2007; Roelofs *et al.*, 2008; Navrady *et al.*, 2017), suggesting a possible causal relationship. Moreover, higher levels of neuroticism are suggested to be associated with MDD longitudinally (Farmer *et al.*, 2008). Furthermore, no age effects have been observed however, being female increase the risk for MDD (Navrady *et al.*, 2017). Neuroticism is observed to be a stable trait throughout life, this in contrast to MDD which often presents itself as a

recurrent trait (Conley, 1985; Hardeveld *et al.*, 2013). This suggests that the strong correlation between MDD and neuroticism is driven by the effect neuroticism has on MDD and not reversely. Furthermore, the economic burden of neuroticism for societies is high (Cuijpers *et al.*, 2010). Measured per capita the top 5% of individuals with the highest neuroticism scores costs around $12,362 per year. These numbers drop when neuroticism scores decrease to $8,243 in the 10% highest scorers, and $5,572 in the 25% highest scorers. These costs are a combination of multiple factors such as direct medical (health care service), direct nonmedical (personal) and direct nonmedical (e.g. missed work day) costs. On a population level the cost of neuroticism per 1 million individuals falling in the top 25% of neuroticism levels exceed the cost of common mental disorders by 2.5 times (neuroticism = $1.393 billion against $585 million for common mental disorders) (Cuijpers *et al.*, 2010).

There is robust evidence for a significant amount of genetic variation contributing to the trait variance (H^2) ranging from 30% to 50% in twin based studies (Kendler *et al.*, 2006a; Vukasović and Bratko, 2015) and 15% using SNP based heritability (Smith *et al.*, 2016). GWAS yielded solely 11 significantly associated loci (Okbay *et al.*, 2016; Smith *et al.*, 2016) until the large UK Biobank cohort (*n*=329,821) was used. Using this large cohort a SNP-based heritability of neuroticism was estimated at 0.108 (SE=0.005) and 116 loci were detected to influence neuroticism scores with the neuroticism polygenic score explain 2.75% of the variance in neuroticism (Luciano *et al.*, 2018). Especially in neuroticism the low amount of variation explained by GWAS and polygenic scores was expected as dimensions of personality are likely to have a considerable amount of variation attributable to non-additive gene-gene interaction or epistatic effects (Jang, Livesley and Vemon, 1996).

1.6 Genetic epidemiology of Cognition

1.6.1 Heritability, familial aggregation, twin studies, familial risk

For general cognitive ability, the substantial heritability of ‘g’ has been documented in dozens of family, twin and adoption studies (Deary, Johnson and Houlihan, 2009). Haworth *et al.* (2010) performed the largest heritability meta-analysis of general

cognition in 11,000 twins. Interestingly, this analysis is a longitudinal study allowing investigation into if heritability of general cognition changes over time. Haworth *et al.* (2010) observed that not only does general cognition have a high heritability of ~40% in childhood but increases as participants get older to ~55% in adolescence and ~65% in young adulthood (Figure 1.11).

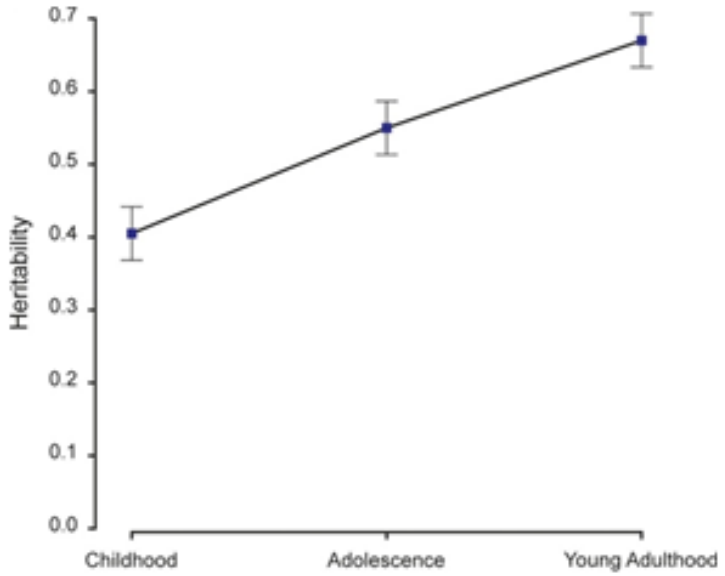


Figure 1.11: Heritability of general cognitive ability in twins from childhood to young adulthood.

Showing linear increased heritability of general cognitive ability in twins from childhood to young adulthood by 25%. Error bars represent ±1 s.e. (bootstrapped with 10,000 samples). Image adapted from Haworth et al. (2010)

Plomin *et al.* (1997) assessed the heritability of cognitive ability using an adoption study from 1 to 16 years of age. The study contains 245 biological mothers who handed over their children for adoption after birth, adoptive parents, the adopted children and 245 control (nonadoptive) parents and their children. Principal components were extracted from specific cognitive ability tests e.g. verbal, spatial, processing speed and memory abilities to assess the relationship of general cognitive ability of biological and adoptive parents and their children at age 3, 4, 7, 9, 10, 12, 14 and 16 years. Figure 1.12 shows the correlation of general cognitive ability between biological parents and adopted away children, adoptive parents and their

adopted children and between biological parents and their biological non-adopted away children (controls).

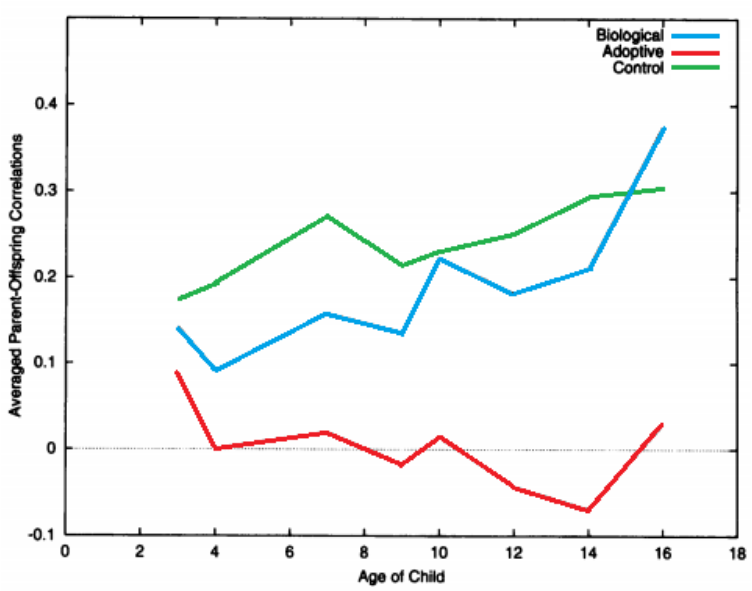


Figure 1.12: Parent offspring correlation of general cognitive ability.

Parent offspring correlation of general cognitive ability between biological parents and adopted away children (blue), adoptive parents and adopted children (red) and biological parents and their biological non-adopted away children (green) at age 3, 4, 7, 9, 10, 12, 14 and 16 years. Image adopted from Plomin et al. (1997).

The results suggest an increasing role of genetic and a decreasing role of environmental factors. Plomin *et al.* (1997) indicates that the study design is not completely representative of the US population due to a higher average socioeconomic status of the study group compared to the general population and higher average performance on the cognitive battery.

Bouchard and McGue. (1981) was one of the earliest large familial aggregation review studies of intelligence. Bouchard and McGue. (1981) used 111 studies that reported on familial resemblance in broad cognitive ability. The study observed that the more strongly individuals are biologically related the higher the average correlation between their general cognition (Figure 1.13).

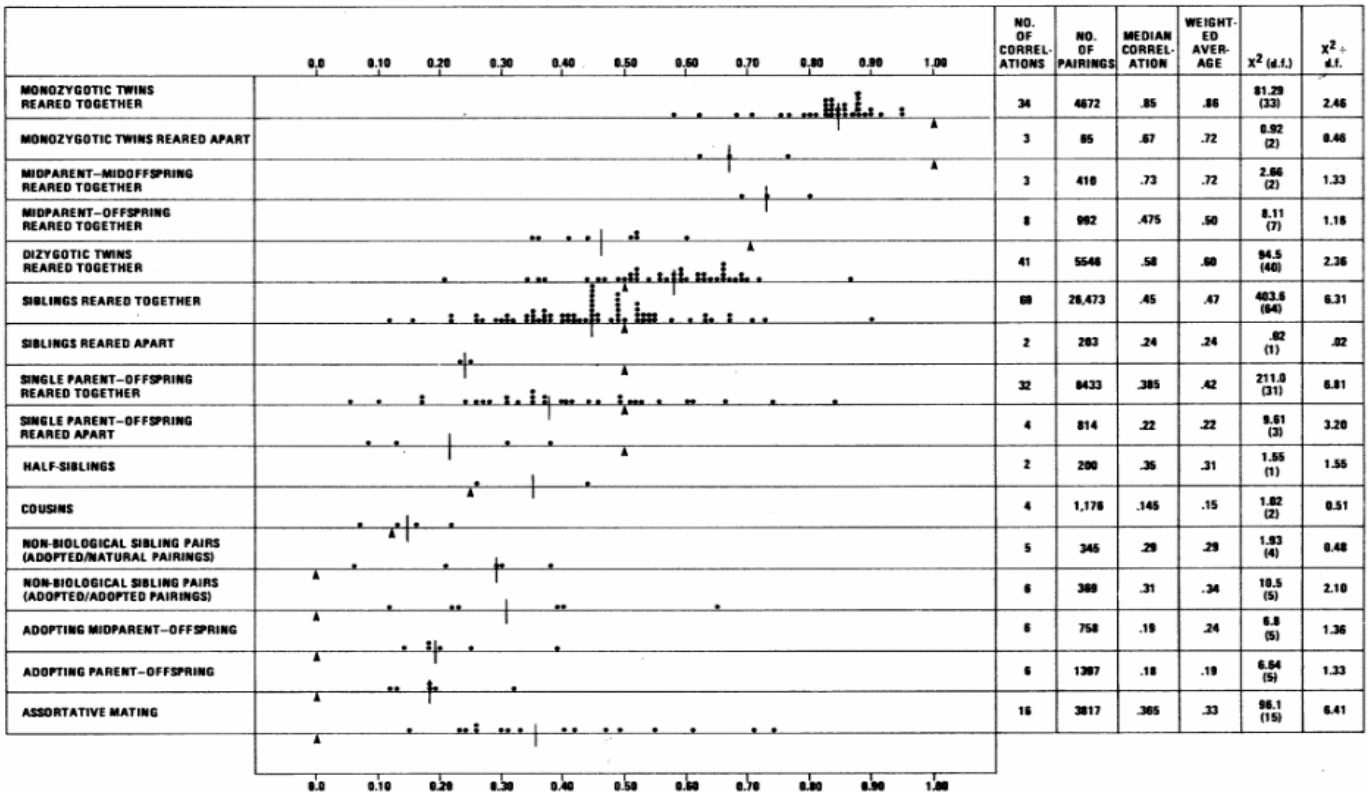


Figure 1.13: Familial correlations for IQ.

Median correlations are indicated by means of a vertical bar and the arrow is the correlation predicted by a simple polygenic model. Image from Bouchard and McGue (1981)

Luciano *et al.* (2010) assessed the heritability of a diverse cognitive battery in a large, pedigree-based cross-sectional study of Scottish families (*n*=1983 families; 6086 individuals). Luciano *et al.* (2010) estimated the heritability of cognitive performance in a range depending on the cognitive domain studied i.e. the heritability of digit symbol performance was the lowest with 36%, followed by verbal fluency and logical memory (both 40%), general cognitive ability 'g' (43%) and finally Mill Hill vocabulary (language; 53%).

1.6.2 Linkage studies

Two related genome-wide family linkage studies of intelligence have been performed (Posthuma *et al.*, 2005; Luciano *et al.*, 2006). The Australian dataset used by Luciano *et al.* (2006) comprised 320 dizygotic twin families (48 families with one non-twin sibling, 10 families with two non-twin siblings and 1 family with three non-twin

siblings) and 41 monozygotic twin families (39 with one non-twin sibling, 2 with two non-twin siblings) showing a heritability of cognitive ability between 0.49 and 0.69 depending on the test. The Dutch study comprised 225 individuals from 100 families, which yielded 159 unspecified sibling pairs showing a full scale IQ heritability of 0.86 (Posthuma, De Geus and Boomsma, 2001). Both studies observed significant linkage on chromosome 2 region q and chromosome 6 region p. The observed linkage with chromosome 2 region q was correlated with performance general cognition. It is worth noting that linkage was observed with verbal cognitive performance but this was near negligible. This might suggest that the observed linkage is not with general intelligence but with the spatial processing cognitive domain. Chromosome 6 region p was associated with general cognitive ability with evidence for linkage with verbal cognitive ability. Both regions await replication.

1.6.3 Replicated candidate genes for cognition

Genes associated with cognitive performance possibly include genes associated with dementia, memory, cardiovascular disease and oxidative stress. Using the Scottish Mental Survey 1932, variations in *KLOTHO* and *NICASTRIN* were reported to be likely involved in general intelligence at both age 11 and 79 (Deary, Hamilton, *et al.*, 2005; Deary, Harris, *et al.*, 2005). *COMT* variations have also shown evidence of involvement in executive cognitive ability (Winterer and Goldman, 2003). Finally, *SSADH* was observed to be associated with IQ (Plomin *et al.*, 2004). Note that while many genes have been associated with cognitive performance many are of small effect and have not been replicated. One gene that has been replicated is *APOE*. Replication analysis (meta-analysis of 38 studies containing 20,000 individuals) showed that possessing the E4 allele of *APOE* was associated in older people with poorer performance in general cognitive ability, episodic memory and executive function (Small *et al.*, 2004). The E2 allele in the same gene seems to act in a protective manner and was not associated with cognition in old age but with cognition in youth (Deary *et al.*, 2002). Apart from APOEε4 and cognitive aging there is nothing replicated for cognitive ability.

1.6.4 Large GWAS consortia

The Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) was formed to facilitate genome-wide association study meta-analyses and replication opportunities among multiple large and well-phenotyped cohort studies. Analyses were performed on many versatile phenotypes which included three different cognitive phenotypes e.g. general cognitive performance 'g' (Davies *et al.*, 2015) , memory (Debette *et al.*, 2015) and executive function and/or processing speed (Ibrahim-Verbaas *et al.*, 2016) using only individuals ≥45 years and of European descent. Davies *et al.* (2015) included 31 population based cohorts participating in the CHARGE consortium (*n*=53,949). SNP based meta-analysis identified 13 genome wide significant SNPs associated with general cognitive performance. All 13 SNPs were located in three genomic regions. The top SNPs (lowest p-value) in each region and genes in these regions were rs10457441 in *MIR2113*, rs17522122 in *AKAP6/NPAS3* and rs10119 in *TOMM40/APOE*. (Debette *et al.*, 2015) included 19 cohorts from the CHARGE consortium comprising 29,076 dementia and stroke free individuals. Two genome-wide statistically significant SNPs were observed associated with poorer delayed recall performance (rs4420638 in paragraph delayed recall and rs13358049 in California verbal learning delayed recall). Rs4420638 was reported to be in LD with APOEε4 which is associated with an increased risk of Alzheimer's disease. Replication analysis (*n*=10,617) was performed for both SNPs with only rs4420638 being significantly replicated. Rs4420638 explained around 1% of variance of paragraph delayed recall performance in the combined dataset (discovery + replication datasets). Ibrahim-Verbaas *et al.* (2016) included 20 cohorts contributing one or more cognitive tests measuring executive function and/or processing speed total. One SNP reached genome wide significance in the meta-analysis discovery GWAS (*n*=5,429–32,070) and in the joint discovery and replication (*n*=1311-21860) meta-analysis GWAS for processing speed (Letter-Digit Substitution and Digit-Symbol Substitution Test). Intronic variant rs17518584 located in gene *CADM2* on chromosome 3. This gene is also known as *SYNCAM2*.

Davies *et al.* (2016) performed a non-meta-analysis based GWAS of reasoning (verbal-numeric reasoning), processing speed (reaction time) and memory (declarative memory) in the UK Biobank sample (*n*=112,151). For verbal-numeric reasoning 149 SNPs from three genomic regions were observed to be significantly associated. Genes in these regions include *CYP2D6*, *NADH*, *NDUFA6*, *SEPT3*, *PDE1C* and *FUT2*. For reaction time 36 SNPs from two genomic regions surpassed the genome-wide significance threshold, including *SH2B3* and *SPATS2L*. Davies *et al.* (2016 observed no genome-wide significant SNPs associated with memory scores.

Sniekers *et al.* (2017) combined 13 cohorts e.g. UK Biobank and the Childhood Intelligence Consortium and performed a GWAS of intelligence (spearman's g or fluid intelligence) on unrelated individuals (*n*=78,308). The meta-analysis identified 336 SNPs located in 18 independent genomic regions surpassing the genome-wide significance threshold. Around half of all associated SNPs are intronic (162/336). The top SNPs fall in 22 genes of which 11 were not previously observed.

1.7 Review of machine learning methods

All methods described until this point are proven to be effective but are also relatively simplistic and still fail to explain the full biological underpinning. In this section numerous methods are described that either a.) Model epistatic interaction using non-parametric decision trees therefore not limiting themselves to a fixed interaction size and b.) Machine learning methods that perform regression/prediction analyses using some form of feature selection on large genetic datasets this will provide a model containing the most informative features. This approach will potentially give a better insight into the complex genetic architecture of human traits compared to e.g. linear regression analysis.

1.7.1 Least Absolute Selection and Shrinkage Operator and Elastic net regularization

Least Absolute Selection and Shrinkage Operator (LASSO) (Tibshirani, 1996) is a coefficient-shrunken version of ordinary least square (OLS; Equation 1.4) estimated by limiting the sum of the absolute value of coefficients not be larger than a constant value (Equation 1.5). This idea is commonly referred to as 'penalised regression'.

All subsequent equations are derived from Tibshirani (1996).

$$OLS = min\ SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (1.4)$$

Where y=observed outcome and ŷ=predicted outcome for individual *i*.

$$(\hat{\propto}, \hat{\beta}) = \arg min \left\{\sum_{i=1}^{N}\left(y_i - \propto - \sum_j \beta_j x_{ij}\right)^2\right\};\ subject\ to\ \sum_j |\beta_j| \leq t \quad (1.5)$$

Where y=observed outcome, α=alpha (0=Ridge regression, 1=LASSO and 0>α<1=Elastic net), β=coefficient of predictor *j*, x=value of predictor *j* in individuals *i* and t=constant value.

LASSO balances two different ideas:

1. Fitting a model between the outcome and the predictors
2. Shrinking the β_s of these predictors.

A penalty parameter λ is imposed on the β_s of predictors that do not improve the model. This can be added to the equation (Equation 1.6) and rewritten to (Equation 1.7):

$$(\hat{\propto}, \hat{\beta}) = \arg min \left\{ \sum_{i=1}^{N} \left(y_i - \propto - \sum_j \beta_j x_{ij} \right)^2 \right\} + \lambda \sum_{j=1}^{p} |\beta_j| \quad (1.6)$$

Where y=observed outcome, α=alpha (0=Ridge regression, 1=LASSO and 0>α<1=Elastic net), β=coefficient of predictor *j*, x=value of predictor *j* in individuals *i*, λ=penalty parameter and t=constant value.

$$\hat{\beta}^{lasso} = \arg min \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \quad (1.7)$$

Since the overall magnitude of the coefficients is limited to not be larger than a constant value *t*, as λ increases (λ = 0; Ordinary Least Squares) important predictors are included in the model, while less important predictors are shrunk, potentially to 0 and therefore excluded. The LASSO solution is the first place where the contour touches the constraint region (Figure 1.14).

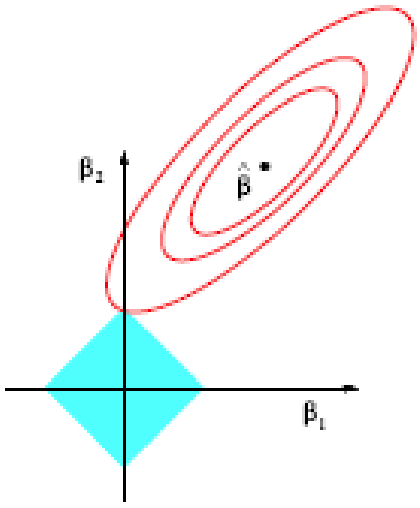


Figure 1.14: Estimation picture of LASSO using 2 predictors.

The solid blue area represents the constraint region $|\beta_1| + |\beta_2| \leq t$. *The red ellipses are the contours of the least squares error function. Image taken from Tibshirani (1996)*

To circumvent the problem of overfitting, the optimal λ setting will be determined by means of a training set and subsequently applied to the test set. Notable disadvantages of the LASSO are that it selects at most *n* variables. The number of selected

predictors is bounded by the number of samples; this becomes problematic when *p* outnumbers *n* (Zou and Hastie, 2005). Also, LASSO fails to do grouped selection; in other words, it tends to select one variable from highly correlated "grouped" variables and ignore the others.

To counter Ridge Regressions and LASSOs shortcomings Zou and Hastie (2005) introduced elastic net. The l_1 penalty ($\lambda \sum_{j=1}^{p} |\beta_j|$) applied by LASSO generates a sparse model and the l_2 penalty ($\lambda \sum_{j=1}^{p} {\beta_j}^2$) never reduces a predictor to 0. Therefore Zou and Hastie (2005) proposed adding a quadratic penalty to the LASSO formula (Equation 1.8).

$$(\hat{\propto}, \hat{\beta}) = \arg min \|y - X\beta\|^2 + \boldsymbol{\lambda}\|\boldsymbol{\beta}\|_{\mathbf{1}} + \boldsymbol{\lambda}\|\boldsymbol{\beta}\|^{\mathbf{2}} \quad (1.8)$$

By adding a quadratic penalty the new method:

- removes the limitation on the number of selected variables
- allow the selection of grouped predictors
- stabilises the l_1 regularisation path (less conservative)

Due to the mixture of a quadratic and non-quadratic penalty the constraint region of elastic net falls between the diamond shaped constraint region of LASSO (non-quadratic) and the circular shaped constraint region of ridge regression (quadratic) (Figure 1.15).

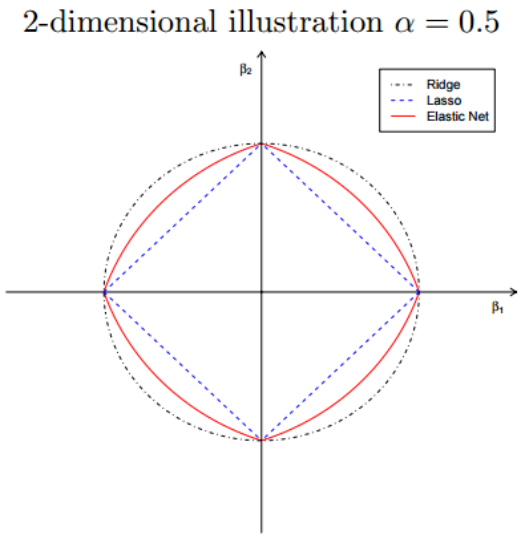
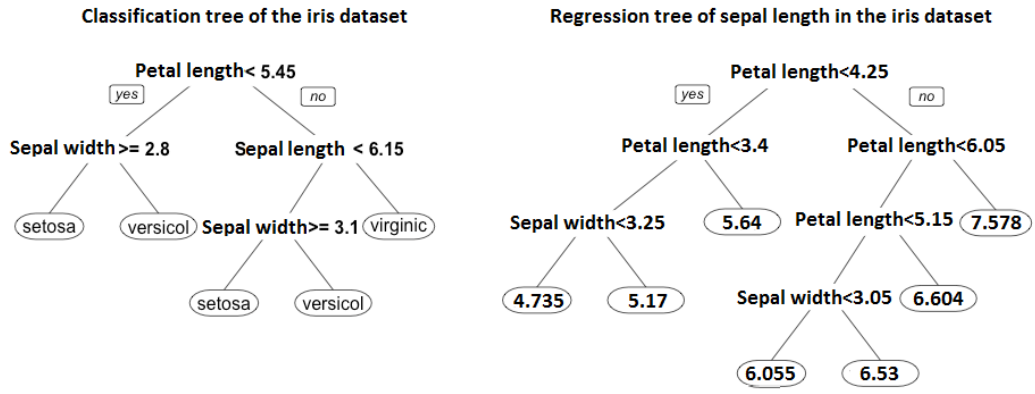


Figure 1.15: Difference in constraint region shapes of penalty scores between LASSO, Ridge regression and Elastic net.

l_1= LASSO regression (blue dotted line), l_2 = Ridge regression (dark blue dotted line) and $l_1 + l_2$=Elastic net (red solid line) α=0.5. Image from Zou and Hastie (2005)

1.7.2 Classification and Regression Trees (CART)

CART are decision tree based methods that can be interpreted as a set of "questions" that lead along a path to a final prediction. CART methods try to utilise the right classifiers (measurements) to "split" the data into partitions. The difference between a classification and regression tree is very straightforward, in that a classification tree is used to predict or explain responses on a categorical dependent variable (Figure 1.16A) while regression trees are used to predict or explain responses on a continuous dependent variable (Figure 1.16B).



A **B**

Figure 1.16: Classification and regression tree on the Iris dataset.

Simple representation of a classification tree (A) and regression tree (B) of the Iris dataset. The Iris data set is a well-known dataset consisting of 50 samples from each of the three species of Iris flowers (Iris setosa, Iris virginica and Iris versicolor). Of these three species the length and the width of the sepals and petals in centimetres were measured.

CART methods solely use the Boolean operator OR to make split a classifier i.e. petal length ≤ 5.45 *OR* petal length > 5.45 (Figure 1.16B). CART methods will grow a tree by including classifiers (recursive partitioning) calculating for every split the 'impurity' or misclassification rate and will define a split with the lowest impurity. Multiple impurity measurements are commonly used i.e. entropy (Equation 1.9) and the Gini index (Equation 1.10) for the classification based methods and sum of square residuals (Equation 1.11) for regression based methods.

$$Entropy;\ H(X_m) = -\sum_k \hat{p}_{mk} \log(\hat{p}_{mk}) \quad (1.9)$$

$$Gini\ Index: H(X_m) = \sum_k \hat{p}_{mk}\left(1 - \hat{p}_{pk}\right) \quad (1.10)$$

$$Sum\ Squared_{(Residuals)} = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \quad (1.11)$$

$\hat{p}_{mk}$ is the proportion of observations from class *k* in node *m*.

CART methods will keep recursively partitioning the dataset until no split can be made that decreases the impurity or when the size of the terminal nodes is less than some value or is 1. This will most often lead to a large tree where some terminal node only contains a small number of individuals. The complexity of a tree can be decreased by pruning sections of the tree that provide little power to classify observations.

1.7.3 Logic trees and logic regression

Logic trees are based on the same principle as a CART; however, it attempts to construct decision trees using Boolean operator combinations (AND, OR and NOT) of binary predictors (Figure 1.17) compared to the Boolean operator OR that is solely used the standard CART method.

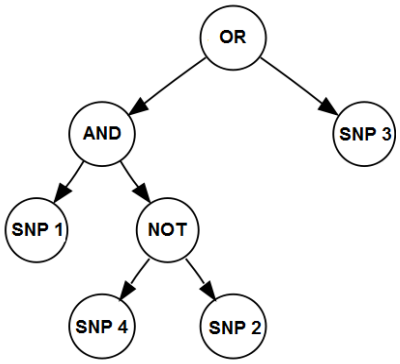


Figure 1.17: Visual representation of a small Logic tree.

Simplified visual representation of a small Logic tree combining all three operators (AND, OR and NOT).

Figure 1.17 can be written out in multiple ways:

Verbally: SNP3 *or* SNP1 *and not* SNP4 *and not* SNP2

Logically: SNP3 ∨ (SNP1 ∧ (*!*SNP4 ∧ *!*SNP2))

Binary: 1 ∨ (1 ∧ (0 ∧ 0))

Logic regression is a non-parametric adaptive regression method developed by Ruczinski, Kooperberg and LeBlanc (2003) which attempts to find Boolean combinations of binary predictors (logic trees) that minimises the scoring function associated with a model type (i.e. residual sum of squares for quantitative outcomes) by estimating the coefficients (βs,) and Boolean expressions (Ls) at the same time (Equation 1.12).

$$Y = \beta_0 + \beta_1 L_1 + \beta_2 L_2 + \ldots\ldots\ldots \beta_p L_p \qquad (1.12)$$

Logic regression applies a greedy hill climb algorithm where the algorithm keeps adding predictors to the model as long as the misclassification rate goes down and only stops when the misclassification rate goes up. By doing this logic regression risks (depending on the random starting point) including logic trees that do not necessarily reflect the best Boolean combinations of binary predictors to properly describe the model accurately (global optimum, Figure 1.18 yellow star, local optimum, Figure 1.18 red stars).

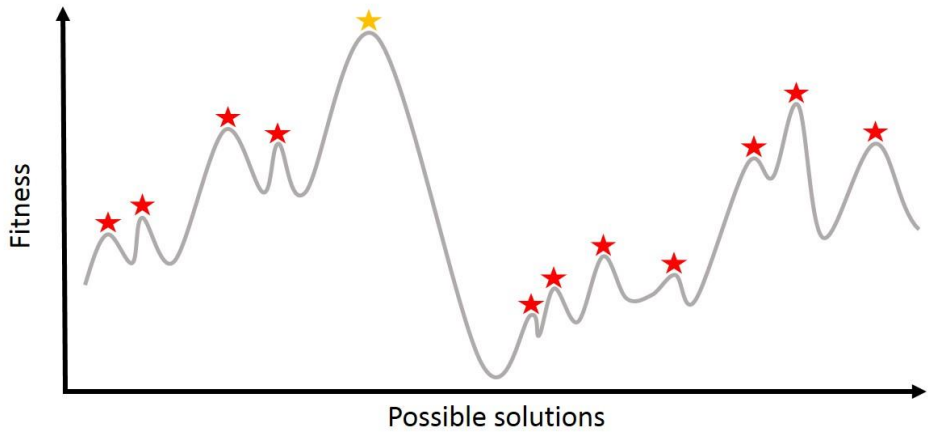


Figure 1.18: Greedy hill climb algorithm in Logic Regression.

Searching for the lowest misclassification error (highest fitness). The objective is to get detect the global optimum (gold star); however, a simple hill climb algorithm, as there are many local maxima (red stars) will possibly select one of those.

Simulated annealing (SA) is used as a search technique to locate a good approximation of a global optimum by allowing misclassification in contrast to the greedy hill climber search. Simulated annealing is performed in 3 steps starting from a randomly selected point:

1. Given a certain state (Boolean combination of genotypes), move to a randomly selected other state in the search space (Figure 1.19).

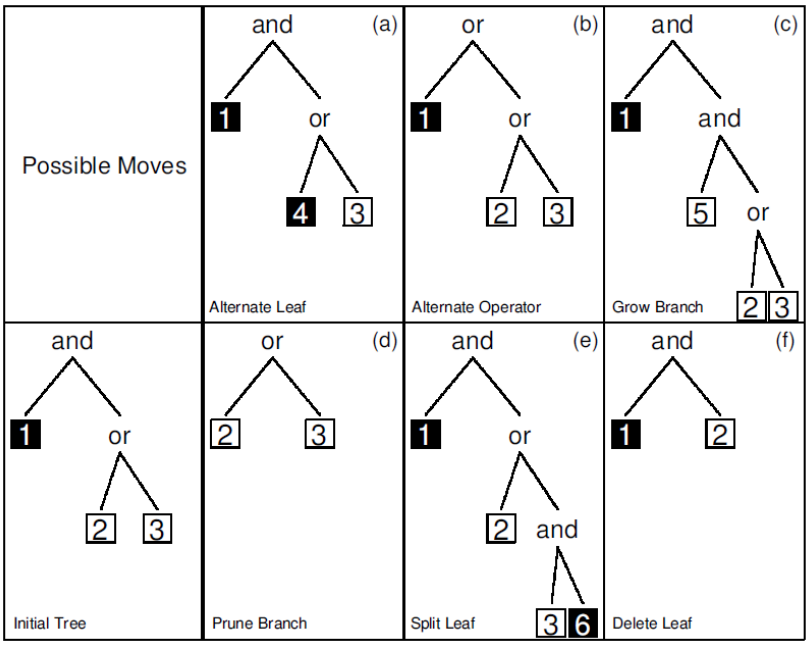


Figure 1.19: Options during Simulated Annealing.

Possible moves considered for each split. Image from Ruczinski, Kooperberg and LeBlanc (2003)

2. Accept the state when the misclassification goes down *or* accept when misclassification rate goes up but stays within predefined acceptance probability.

3. Repeat step 1 and 2 *i* times until either misclassification rates do not decrease anymore or when the predefined acceptance probability is exceeded.

Note that SA starts at a high "temperature" corresponding with a higher acceptance probability and "cools down" for each iteration *i* making the acceptance probability smaller and therefore stricter, leaving less space for allowing misclassification. As mentioned previously, logic trees are constructed using binary predictors. Logic regression has proven to perform well on dichotomised genetic data (Kooperberg *et al.*, 2001; Ruczinski, Kooperberg and LeBlanc, 2004; Schwender, 2007) however applying logic regression on genotype data (homozygote allele 1: 0; heterozygote: 1 and homozygote allele 2: 2) might be more desirable as it represents a more complete working of genetics. Genetic Programming for Association Studies (GPAS) (Nunkesser *et al.*, 2007) combines the logic regression method of using Boolean operators (AND, OR and NOT) but allows the splits to be made on multilayer genotypes (Figure 1.20).

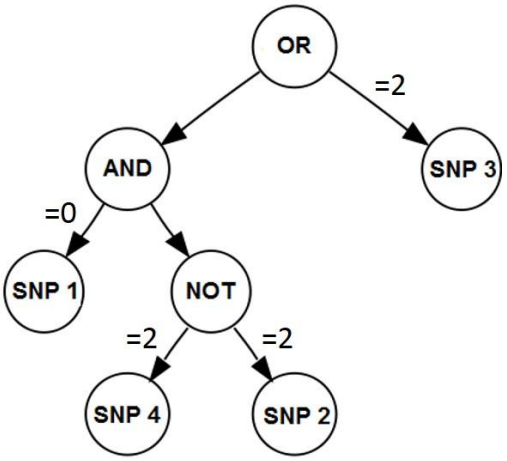


Figure 1.20: Visual representation of a small Logic tree using genotype data.

Simplified visual representation of a small Logic tree combining all three operators (AND, OR and NOT) on genotype data.

1.7.4 C5.0 ruleset

C5.0 is a modified version of Quinlan's C4.5 classification model (Quinlan, 1992). Both C4.5 and C5.0 allow rule-based models and evaluation of variable importance (Wu *et al.*, 2008; Kuhn and Johnson, 2013). C5.0s non-parametric algorithm builds decision trees using information entropy Equation 1.13, at each node splits are chosen using normalised information gain which acts as the purity criterion.

$$I_E(p) = -\sum_{i=1}^{m} p_i \log p_i \quad (1.13)$$

With p_i being the probability of a given class as the outcome for each of *m* possible classes. In short, if the *p* of classes is somewhat balanced the information entropy will be higher. C5.0 calculates the difference in information entropy before and after a split. The information entropy before the split ($info^{S}_{before}$) is calculated by Equation 1.13 above. For split *S* with *K* partitions the information entropy for each resulting partition is calculated and multiplied by the proportion of samples assigned to that partition (n_i) given the total number of samples (*n*) after the partition. By doing so a weight is given to each partition. This is subsequently summed over all possible partitions *K* to give the information entropy after split *S* Equation 1.14.

$$info^{S}_{after} = \sum_{i=1}^{k} info_i \frac{n_i}{n} \quad (1.14)$$

The information gain of split *S* can now be calculated by $info^{S}_{before}$ - $info^{S}_{after}$. A positive information gain implies a lower information entropy after the split than before, therefore the uncertainty decreased. The information gain is normalised to allow for the consideration of each class. The class with the highest normalised information gain is selected. This process is repeated for smaller subsets. Branches that do not contribute to the improvement of the trees classification are removed using rule-based pessimistic pruning. During rule-based pessimistic pruning each top to bottom path from the initial tree is collapsed into a rule. C5.0 evaluates each rule on independent conditional statements, thereby assessing whether or not they can be generalised by removing terms in the conditional statement. Error rates created in this process are compared to a pre-determined pessimistic error rate. If a rule passes the pessimistic error rate it is removed, when none is above the error rate the worst performing rule is removed. The pessimistic error rate is recalculated and the process is repeated on the ever shrinking tree until all conditions are above the baseline error

rate or all removed leading to the removal of the rule. Finally, C5.0 calls a vote for each rule for the most likely class and the class with the highest vote is used. This results in a pruned tree where each possible combination from the top node to bottom node (Figure 1.21) in the tree is a so called *ruleset* and can be used as a predictor in a regression model.

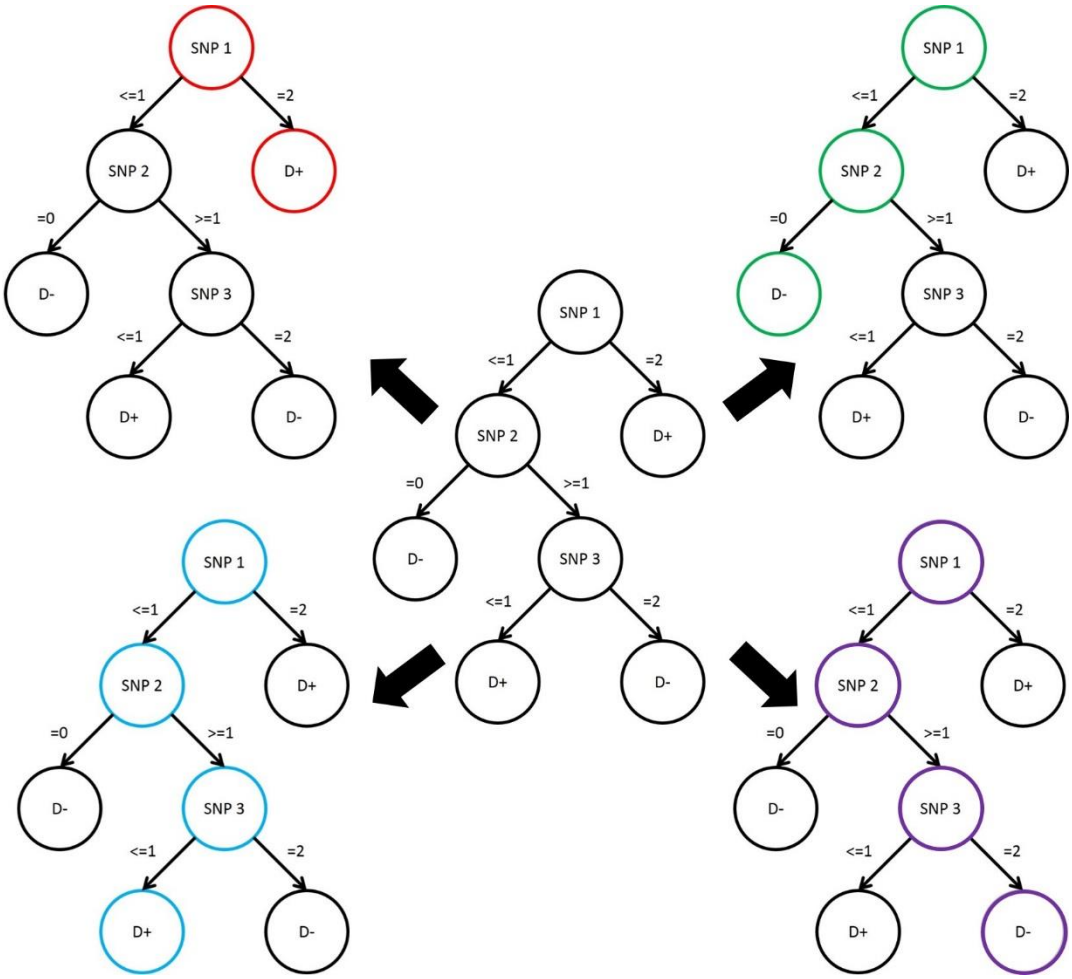


Figure 1.21: C5.0 example tree.

The middle tree represents the original tree; every colour path indicates a ruleset. D+ = disease positive and D- = disease negative.

1.8 Overall hypothesis and aims of this thesis

The aim of this thesis can be split into seemingly unrelated sections:

1) Identifying cognitive performance differences between MDD cases and controls as well as between single-episode and recurrent MDD and linking these differences to genetic components using standard methodologies (e.g. GWAS and PGRS)

2) Assessing the capabilities of two tree based method to detect epistasis under simulated circumstances even in the presence of a substantial polygenic component.

3) Applying all methodologies described in point 1 and 2 separately on a trait of interest. Results of the methodologies are combined and analysed as a whole using a machine learning method.

I hypothesise that by doing this "agnostic" methodology a larger proportion of variation might be explained compared to using standard methodologies.

Chapter 2

Phenotypic and genetic analysis of cognitive performance in Major Depressive Disorder in the Generation Scotland: Scottish Family Health Study

Joeri J. Meijsen, Archie Campbell, Caroline Hayward, David J. Porteous, Ian J. Deary, Riccardo E. Marioni and Kristin K. Nicodemus

Work presented in this chapter has been peer reviewed and published in

Translational Psychiatry 2018 vol: 8 (1) article: 63

doi: 10.1038/s41398-018-0111-0

Brief introduction

This chapter outlines published work on the investigation of phenotypic differences in cognitive ability between controls and individuals with single-episode and/or recurrent Major Depressive Disorder. For observed cognitive ability differences standard and proven statistical methodologies such as genome-wide single locus, genome-wide single locus interaction, polygenic and polygenic interaction analyses were performed to detect single locus and polygenic component associations.

Statement outlining the contribution of first author and co-authors

All analyses and writing of this paper have been done by Joeri Meijsen. Co-authors mentioned on this paper have read, edited and approved the final version sent for submission.

2 Phenotypic and genetic analysis of cognitive performance in Major Depressive Disorder in the Generation Scotland: Scottish Family Health Study

2.1 Abstract

Lower performances in cognitive ability in individuals with Major Depressive Disorder (MDD) have been observed on multiple occasions. Understanding cognitive performance in MDD could provide a wider insight in the aetiology of MDD as a whole. Using a large, well characterised cohort (*n*=7012), we tested for: differences in cognitive performance by MDD status and a gene (single SNP or polygenic score) by MDD interaction effect on cognitive performance. Linear regression was used to assess the association between cognitive performance and MDD status in a case-control, single episode-recurrent MDD and control-recurrent MDD study design. Test scores on verbal declarative memory, executive functioning, vocabulary, and processing speed were examined. Cognitive performance measures showing a significant difference between groups were subsequently analysed for genetic associations. Those with recurrent MDD have lower processing speed versus controls and single-episode MDD (β = -2.44, p=3.6x10^{-04}; β = -2.86, p=1.8x10^{-03}, respectively). There were significantly higher vocabulary scores in MDD cases versus controls (β=0.79, p=2.0x10^{-06}**),** and for recurrent MDD versus controls (β=0.95, p=5.8x10^{-05}**)**. Observed differences could not be linked to significant single-locus associations. Polygenic scores created from a processing speed meta-analysis GWAS explained 1% of variation in processing speed performance in the single episode versus recurrent MDD study (p=1.7x10^{-03}) and 0.5% of variation in the control versus recurrent MDD study (p=1.6x10^{-10}). Individuals with recurrent MDD showed lower processing speed and executive function while showing higher vocabulary performance. Within MDD, persons with recurrent episodes show lower processing speed and executive function scores relative to individuals experiencing a single episode.

2.2 Introduction

Major Depressive Disorder (MDD) is common mental disorder affecting at least 1 in 10 in the UK (D. J. Smith *et al.*, 2013) and is a leading cause of disability worldwide. Showing a SNP-based heritability of ~30% (Lubke *et al.*, 2012; Fernandez-Pujals *et al.*, 2015) and a twin-based estimate of ~40% (Kendler *et al.*, 2006), MDD has a substantial genetic component. It has been shown that individuals suffering from MDD show lower performance in cognitive domains such as executive function (EF), memory, language and attention (Lim *et al.*, 2013; Snyder, 2013; Cullen *et al.*, 2015). The identification and quantification of lower cognitive performance in MDD could lead to a better understanding of the underlying aetiology of depression, to improve treatment of patients, or as an endophenotype for subsequent studies investigating the genetic architecture of MDD. These targeted approaches could possibly lay the groundwork to improve the mental health of MDD patients and therefore lower the burden MDD has on society.

Despite the high prevalence of MDD, cognitive lower scores in MDD have not been as widely studied as other psychiatric disorders such as bipolar disorder (Bora and Pantelis, 2015) and schizophrenia (Bowie and Harvey, 2006; Bora and Pantelis, 2015). Snyder et al. (Snyder, 2013) performed an extensive and the largest-to-date meta-analysis of cognitive performance in MDD, focussing mainly on tasks that measure executive function (EF) with the exception of two non-EF tests measuring vocabulary (language) and digit symbol substitution (processing speed, but is also considered by some to be a component of EF). They observed that MDD patients showed a lower performance in phonemic verbal fluency and digit-symbol measures. That is, MDD patients produced significantly fewer words than healthy control individuals and recoded significantly fewer symbols to digits in digit symbol measures. Vocabulary performance was observed to be lower in MDD patients; however, the effect was not significant. Logical memory (LM) immediate and delayed (both measuring verbal declarative memory) have been less well-studied compared to other cognitive measures in depression. Lim et al. (2013) (Lim *et al.*, 2013) conducted the largest meta-analysis study of LM to date (*n* logical memory immediate=291; *n* logical memory delayed=348). They observed that MDD patients performed significantly less

well than controls on both LM immediate and delayed. This result has been previously reported by smaller studies not included in the Lim et al. study (Delgado, Kapczinski and Chaves, 2012; Maeshima *et al.*, 2013), with one exception (Travis *et al.*, 2014). Significant lower performances were also observed in the attention domain (Lim *et al.*, 2013), via the digit span test and continuous performance test where MDD patients performed slower compared to controls. The final domain examined, visuospatial processing (immediate and delayed visual memory), showed no differences between MDD patients and controls (Lim *et al.*, 2013).

As the genomic underpinnings of MDD are poorly understood (Ripke *et al.*, 2013), we examined genomic associations with cognitive differences as observed in our study as an endophenotype strategy. Using the extensively phenotyped Generation Scotland Cohort Study, we sought to: (a) investigate whether cognitive ability in MDD patients differs from controls without MDD or reported mental illness, (b) assess whether cognitive performance differs between single-episode MDD versus recurrent MDD, (c) investigate cognitive performance between controls and recurrent MDD and (d) to reduce multiple testing we performed genome-wide single locus, genome-wide single locus interaction, polygenic and polygenic interaction analyses only on cognitive performance tests showing a significant difference within study designs. This study represents the largest single cohort study investigating the association of cognitive performance in depression using a formal clinical diagnosis of MDD and incorporating genomic association analyses. The largest single cohort study investigating cognitive performance in depression is the UK Biobank cohort study (Cullen *et al.*, 2015) however that study relies on self-reported MDD status and does not examine genetic factors.

2.3 Materials and Methods

2.3.1 Cohort data and phenotyping

Generation Scotland: the Scottish Family Health Study (GS:SFHS) is a family-based cohort study sampled from the general population in Scotland (www.generationscotland.org) (Smith *et al.*, 2006; B. H. Smith *et al.*, 2013). The

study design has been widely documented (Smith *et al.*, 2006; B. H. Smith *et al.*, 2013). In short, between 2006 and 2011 over 24,000 subjects were recruited into the study. The initial sample of study subjects (*n*=7,953) were registered with general medical practitioners, between 35 and 65 years, and from five regions of Scotland. These initial study subjects were asked to bring a relative within the age range 18-99 to the baseline data collection. Participants were asked to fill in health, lifestyle and family history questionnaires and answer a 30 minute interview which included questions about possible mental ill health. If participants answered positively on either of the 2 mental health screening questions ("*Have you ever seen anybody for emotional or psychiatric problems?*" and "*Was there ever a time when you, or someone else, thought you should see someone because of the way you were feeling or acting?*") (*n*=4,539), they were asked to take part in a Structured Clinical Interview for DSM-IV (SCID) (First *et al.*, 1997), focussing on mood disorders. Individuals answering "no" to both questions were assigned to the control group. Individuals who completed the SCID but did not meet the criteria for MDD or bipolar disorder were subsequently assigned to the control group (Clarke *et al.*, 2015) (*n*=1,727). Finally, individuals who were invited for the SCID interview but refused to take part (*n*=507) were not assigned a either case or control group (Fernandez-Pujals *et al.*, 2015).

Four cognitive domains were measured in Generation Scotland: processing speed (Wechsler Digit Symbol Substitution Test; recoding symbols to digits (Wechsler D., 1998a) -DST), verbal declarative memory (Wechsler Logical Memory Test; sum of immediate and delayed recall of an oral story (Wechsler D., 1998b) – LM1 and LM2), executive functioning (the phonemic verbal fluency test; using the letters C, F, and L, each for one minute (Lezak, M.; Howieson, D.; Bigler, E.; Tranel, 2012) - VFT), language (the Mill Hill Vocabulary Scale, Junior and Senior Synonyms combined – finding a synonym of a given word (Raven JC, Court JH, 1977) - MHVS) and the difference between logical memory immediate and delayed (LM1-LM2). The correlation between scores on tests of these different cognitive domains are reported in Supplementary Tables S2.1-S2.4.

In addition to age and sex, we selected lifestyle factors (self-reported smoking and alcohol intake), socioeconomic status (the Scottish Index of Multiple Deprivation (Payne and Abel, 2012)), medication usage (anti-depressants and mood stabilisers) and 15 genetic principal components to control for population stratification. These variables have been previously used as covariates in Cullen et al, 2015 (Cullen *et al.*, 2015) to investigate cognitive differences in depression using the UK Biobank cohort.

2.3.2 Genetic data

DNA of 20,128 GS:SFHS participants was analysed by means of high density genome wide bead array genotyping (Illumina OmniExpress 700K SNP GWAS and 250K exome chip). We selected a set of unrelated individuals for use in our analyses, to remove the influence of shared environments. We removed single nucleotide polymorphisms (SNPs) and individuals with a missingness of >1% and removed rare SNPs with a minor allele frequency < 0.01 leaving 557,292 SNPs for analysis. We used Genome-wide Complex Trait Analysis (Yang *et al.*, 2011) to extract a list of genetically-unrelated individuals from a predefined list of participants with a known MDD SCID diagnosis or controls. Seven thousand one hundred and seventy-two unrelated individuals (relatedness < 0.025, corresponding to second degree cousins) were selected, of which 1,042 individuals (14.5%) were diagnosed with either single or recurrent depression. One hundred and five individuals were removed due to the lack of self-reported medical background information. Another 25 individuals with self-reported Alzheimer's and/or Parkinson's disease were removed leading to a total of 7,012 individuals, of which 1,021 individuals (14.5%) were diagnosed with a form of depression.

2.3.3 Statistical analysis – phenotypic differences

We used phi coefficients and Spearman correlation coefficients to determine the level of correlation between the pool of potential covariates and MDD case-control, single-recurrent or control-recurrent status. As a continuous variable, age was assessed

using the Spearman correlation coefficient. As all other variables were binary, their correlations were assessed using the phi coefficient, with associated *p*-values from either a χ^2 or Fisher's exact test. The Fisher's exact test was used when observed cell counts in the 2 x 2 contingency table were less than 5. No potential covariate was strongly correlated with MDD case-control (Supplementary Table S2.5), single-recurrent (Supplementary Table S2.6) or control-recurrent (Supplementary Table S2.7) status aside from age, sex and medication usage in the case-control study and solely medication usage in both the single-recurrent and control-recurrent MDD study, as expected. To keep in line with Cullen et al, 2015 all covariates (sex, age, alcohol consumption, smoking tobacco, medication usage, socioeconomic status and 15 principal components) were included in the full model.

Multiple regression analysis was performed for each cognitive test and the diagnosis label before and after controlling for covariates. We used the following models: a baseline model (Equation 2.1):

$$Cognitive\ ability\ test_k = \beta_0 + \beta_{diagnosis\ label} diagnosis\ label \quad (2.1)$$

and a full model (Equation 2.2):

$$Cognitive\ ability\ test_k = \beta_0 + \beta_{diagnosis\ label} diagnosis\ label + \sum_{i=1}^{n} \beta_i Covariates_i \quad (2.2)$$

We observed that medication usage contained many missing values (52%), with only a small percentage of all participants answering positively (5.1%). Therefore, we performed model 2 and all subsequent analyses twice 1.) including medication usage (Figure 2.1 - M2A) and 2.) excluding medication usage (Figure 2.1 - M2B) as a

covariate. A Bonferroni significance level of $p < 8.3\times10^{-03}$ (p = 0.05/6 cognitive tests) was used. All models were run using the R Statistical Computing Environment (R Core Team, 2017) v 3.1.0.

2.3.4 Statistical analysis – Single Locus analysis

We performed a Genome-Wide Association Study (GWAS) for the cognitive performance variables that showed a significant difference in the phenotypic analyses. We further tested whether each SNP's association with cognitive performance depended on MDD status via a Genome-Wide by Environment Interaction Study (GWEIS). The GWAS analyses can be seen as a baseline model and GWEIS as a measure of non-additive effects for SNP and depression case status. The standard Bonferroni significance level of $p < 5\times10^{-08}$ is conservative, as many SNPs are in linkage disequilibrium thus statistical tests are not independent. Therefore, we applied a less conservative significant level $p < 1.52\times10^{-07}$ derived from the Genetic type 1 Error Calculator (GEC) (Li *et al.*, 2012). All models were run using PLINK version v1.90b1g.

2.3.5 Statistical analysis – Polygenic analysis

Polygenic Risk Scores (PGRS) were calculated for five p-value threshold ranges (0–0.01, 0- 0.05, 0-0.1, 0-0.5, 0-1) using summary output from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) meta-analysis GWAS of DST and similar tests that controlled for sex, age, assessment centre, education and community (Ibrahim-Verbaas *et al.*, 2016). Generation Scotland is a part of the CHARGE consortium but was not included in this specific meta-GWAS study. The CHARGE consortium performed a sample size weighted meta-analysis because of the differences in the test methodology and measurement units. The z-statistic was weighted by the effective sample size (sample size × (observed dosage variance/expected dosage variance)) for each SNP. We pruned the Generation Scotland dataset for linkage disequilibrium (window size = 50kb, step size = 5kb and r^2 threshold = 0.25) and converted the CHARGE z-statistics to standardised beta

coefficients using the z-score and standard error provided by CHARGE. We performed a linear regression model between the DST and the polygenic risk scores as well as a model including polygenic risk score-by-MDD status interaction in a controls-recurrent MDD and single-recurrent MDD study. Consistently with previous analyses, we restricted our polygenic score analysis to the groups where we had observed significant differences. We controlled for all covariates and the number of valid genotypes in a model that did not include medication usage. Figure 2.1 shows a graphic representation of performed analyses.

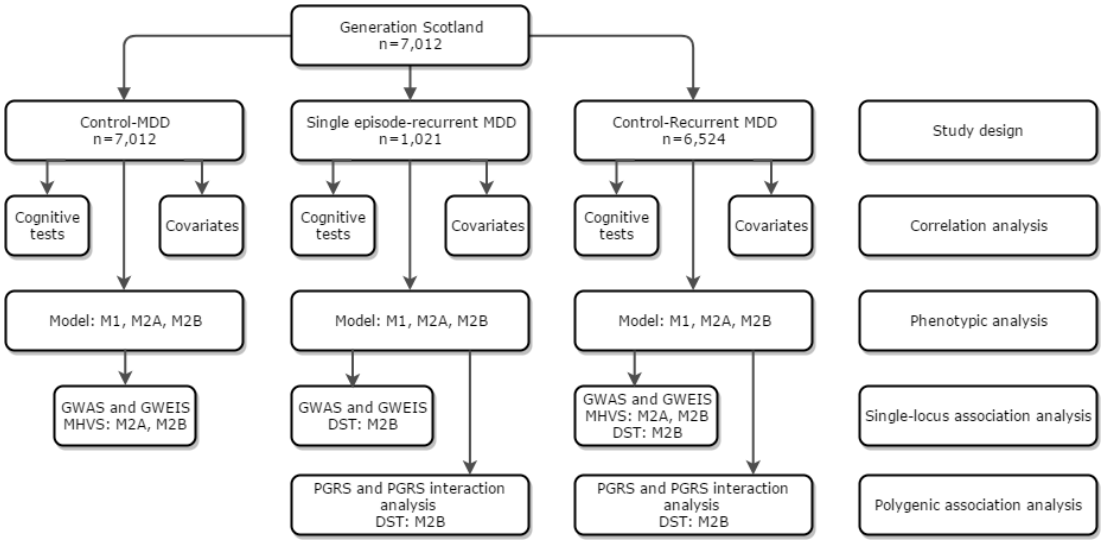


Figure 2.1: Flow chart of performed analyses.

M1 = no covariates, M2A = controlling for all covariates and M2B= controlling for all covariates except medication

2.4 Results

2.4.1 Descriptive statistics

We observed significant differences in the distributions of sex, age, alcohol consumption, smoking tobacco, medication usage and socioeconomic status across MDD status with a higher frequency of females (69-72%), tobacco smokers (23-26.8%) and medication users in the MDD case group (Table 2.1). Within MDD cases, alcohol drinkers represented a significant lower frequency in the recurrent MDD (83.5%) group than the single episode MDD (88.8%) but a higher frequency in

medication usage. On average, controls were slightly but significantly older than cases with MDD and lived in less deprived areas.

Covariate	MDD Status		
	Control	Single	Recurrent
	n=5,991	*n*=488	*n*=533
Sex (*n*, % Female, 0 NA)	3252 (54)	336 (69)	384 (72)*
Age (M, SD, 0 NA)	51.7 (13.8)	49.1 (12.6)	50.1 (11.1)*
Alcohol (*n*, % Drinking, 107 NA)	5410 (91.6)	424 (88.8)	437 (83.5)
Smoking (*n*, % Smoking, 90 NA)	837 (14.2)	111 (23)	141 (26.8)*
Medication (*n*, % Using, 3595 NA)	73 (2.49)	31 (12.4)	71 (29.8)
SES (M, SD, 363 NA)	4080 (1819.3)	3836 (1853.1)	3422.7 (1966.4)

Table 2.1: Demographics and medical history by MDD case status.
All association show significant group differences at 0.05, corrected for multiple testing except for single-episode versus recurrent MDD highlighted with.*

2.4.2 Cognitive association by depression status

We performed three linear regression analyses for each cognitive test (the dependent variable). The predictor variable was MDD diagnosis, classified as either control-MDD, single episode-recurrent MDD and control-recurrent MDD. No other covariates were considered in these baseline models (Table 2.2). No significant association was observed between MDD and cognitive test scores, except for digit symbol substitution in the single episode-recurrent comparison (β = -3.41, p = 5.8×10^{-04}), with the recurrent MDD group recode fewer symbols to digit compared to single-episode MDD group.

	Control-MDD			Single-Recurrent MDD			Control-Recurrent MDD		
	β	Pr(>|t|)	N	β	Pr(>|t|)	N	β	Pr(>|t|)	N
LM1*	0.20	0.12	6974	-0.56	0.01	1021	-0.06	0.72	6486
LM2	0.21	0.13	6936	-0.52	0.03	1016	-0.03	0.86	6452
LM1-LM2	-0.03	0.59	6936	-0.02	0.86	1016	-0.04	0.61	6452
DTS	-0.02	0.96	6936	**-3.41**	**5.7E-04**	**1011**	-1.65	0.02	6452
VFT	0.03	0.93	6934	-0.03	0.96	1019	0.01	0.97	6447
MHVS	0.05	0.75	6887	0.26	0.37	1013	0.17	0.41	6401

Table 2.2: Association between diagnosis label and cognitive performance excluding covariates.

*Association between diagnosis label and cognitive performance in both study designs, without controlling for covariates. *LM1: Logical memory immediate, LM2: logical memory delayed, DST: digit symbol substitution test, VFT: verbal fluency total, MHVS: Mill Hill vocabulary score. Bolded results are significant after Bonferroni correction.*

We then performed linear regression on the full model, including all covariates that were used in Cullen et al (Cullen *et al.*, 2015), which includes medication usage (Supplementary Table S2.8). We observed a significant difference after correcting for multiple testing in the MHVS in both the control-MDD and control-recurrent MDD study. Individuals with depression had higher scores in the MHVS, identifying on average 0.66 more synonyms relative to controls (β = 0.66, $p = 2.96\times10^{-03}$). Between controls and individuals with recurrent MDD, participants with recurrent depression performed even higher, with 1.07 more synonyms identified ($p = 6.0\times10^{-04}$).

When leaving out medication usage (Table 2.3) we observed the same significant higher performance of the MHVS in the MDD and recurrent MDD group in the control-MDD (β = 0.79, $p = 2.02\times10^{-06}$) and control-recurrent MDD (β = 0.95, $p = 5.8\times10^{-05}$) study design. Individuals with recurrent MDD recoded significantly fewer symbols back to digits compared to their study design counterparts in the single episode-recurrent (β = -2.86, $p = 1.8\times10^{-03}$) and control-recurrent MDD (β = -2.44, $p = 3.6\times10^{-04}$) study designs

	Control-MDD			Single-Recurrent MDD			Control-Recurrent MDD		
	β	Pr(>|t|)	N	β	Pr(>|t|)	N	β	Pr(>|t|)	N
LM1*	0.19	0.18	6447	-0.41	0.09	923	-5.0E-03	0.97	6008
LM2	0.15	0.31	6410	-0.36	0.16	918	-0.02	0.88	5975
LM1-LM2	0.01	0.89	6410	-0.03	0.80	918	6.3E-03	0.95	5975
DST	-1.09	0.03	6411	**-2.86**	**1.8E-03**	**913**	**-2.44**	**3.6E-04**	**5976**
VFT	0.89	0.04	6417	0.30	0.69	921	1.04	0.06	5979
MHVS	**0.79**	**2.02E-06**	**6372**	0.42	0.13	916	**0.95**	**5.8E-05**	**5935**

Table 2.3: Association between diagnosis label and cognitive performance including covariates.

*Association between diagnosis label and cognitive performance in both study designs, after controlling for all covariates except medication. *LM1: Logical memory immediate, LM2: logical memory delayed, DST: digit symbol substitution test, VFT: verbal fluency total, MHVS: Mill Hill vocabulary score. Bolded results are significant after Bonferroni correction.*

2.4.3 Single-locus analysis

GWAS (Supplementary Figure S2.1A-B) and GWEIS (Supplementary Figure S2.2A-B) analyses was performed on MHVS in the control-MDD and control-recurrent MDD study designs excluding medication usage. No SNP was observed below the GEC significance threshold in the MHVS analyses (GEC $p = 1.52\times10^{-07}$). The same analysis was performed for DST in the single episode-recurrent and control-recurrent MDD study designs without controlling for medication usage (Supplementary Figures S2.3A-b and S1.4A-B). We did not observe a significant association between genomic variation and DST. Both the strongest non-significant GWAS and GWEIS hit were associated with digit symbol performance and observed in the single episode-recurrent MDD study design. SNP *rs10829637* ($p= 3.3\times10^{-07}$) located on chromosome 10 in *LOC107984280* was the most significant GWAS hit and *rs911684* ($p= 6.7\times10^{-07}$) located on chromosome 14 in *LOC100506999* was the most significant GWEIS hit. Other GWAS and GWEIS results can be found in (Supplementary Figure S2.5A-B, S2.6A-B).

2.4.4 Polygenic score analysis

In the single episode-recurrent study design, the DST PGRS was significantly associated with DST performance at all but two *p*-value thresholds (Bonferroni p =

0.01; 0.05/5 PGRS ranges), indicating that the DST polygenic risk score explained a significant amount of variation (most significant polygenic score: R^2 1%, *p*-value threshold = 0.1, $p = 1.6\times10^{-03}$) in performance among MDD cases (Table 2.4). We observed significant statistical association with each PGRS range in the control-recurrent MDD study design with the PGRS explaining between 0.13 and 0.5% of variation (Table 2.4). However, the effect of the DST polygenic score did not differ between single episode-recurrent cases nor between controls and recurrent MDD cases. We did not observed a PGRS by MDD group interaction on DST performance (Supplementary Table S2.9)

	Single-Recurrent MDD			Control-Recurrent MDD		
Range	Direction	Pr(>|t|)	R^2 (%)	Direction	Pr(>|t|)	R^2 (%)
0-0.01	+	0.14	0.48	+	**1.63E-03**	**0.13**
0-0.05	+	**4.75E-03**	**0.85**	+	**9.95e-06**	**0.23**
0-0.1	+	**1.66E-03**	**1**	+	**5.12e-08**	**0.36**
0-0.5	+	0.011	0.66	+	**7.83e-10**	**0.46**
0-1	+	**8.42E-03**	**0.7**	+	**1.61e-10**	**0.5**

Table 2.4: Association between DST performance and PRS.
Association between DST performance and PRS derived from the DST meta-analysis of the CHARGE consortium. Bolded results are significant after Bonferroni correction.

2.5 Discussion

This study of cognitive performance in MDD is the largest single cohort study with a formal clinical diagnosis of MDD and incorporating genomic association analyse. The only larger single cohort study being UK Biobank, which does not contain a formal clinical diagnosis of MDD and does not investigate genetics associations. Moreover, the cognitive battery used in Generation Scotland is standardised and validated on large representative samples using pre-existing evidence while the cognitive battery used in UK Biobank was bespoke and designed for UK Biobank itself.

We observed significantly higher MHVS scores in MDD cases versus controls, and between recurrent depression versus controls with and without controlling for

medication usage, with 'cases' performing higher than the latter in both studies. The same directionality of effect was observed in UKB by Cullen et al (Cullen *et al.*, 2015); they also observed a significant higher score in vocabulary performance in MDD case groups compared to controls. We also observed significant lower performance of DST between recurrent and single-episode MDD cases, and between recurrent MDD and controls; however, in this case the 'cases' performed less well in both study designs. We also observed a significant amount of variation explained in DST performance using the CHARGE consortium DST polygenic risk score; however, this result was observed across cases and controls and did not differ by case status, indicating that the DST polygenic risk score may not be a useful endophenotype for depression.

Our results are consistent with the largest meta-analysis of case-control differences in digit symbol coding performance, which found that individuals with depression performed significantly lower than controls (Snyder, 2013). One previous study not included in the recent meta-analysis examining differences in digit symbol coding performance between individuals with depression (current (n=37) or previous (n=81)) and controls (n=50) found no significant difference between the three groups, but the sample size was modest (Halvorsen *et al.*, 2012). We also report no significant differences between cases and controls or single-episode versus recurrent MDD on vocabulary, also consistent with (Snyder, 2013). However, we were unable to replicate some results previously reported in the literature (Delgado, Kapczinski and Chaves, 2012; Maeshima *et al.*, 2012, 2013; Gooren, Schlattmann and Neu, 2013; Lim *et al.*, 2013; Snyder, 2013; Travis *et al.*, 2014; Cullen *et al.*, 2015). (Snyder, 2013) observed significant lower performance in phonemic verbal fluency between cases and controls whereas we observed no significant difference. One possible reason is through the inclusion of people in the control group that have symptoms of depression but do not meet the criteria of being diagnosed with MDD, in other words, misclassification of controls, which may have biased our estimates toward the null. Misclassification of controls as MDD participants might be possible due to the screening questions: "*Have you ever seen anybody for emotional or psychiatric*

problems?" and "*Was there ever a time when you, or someone else, thought you should see someone because of the way you were feeling or acting*?". However, this is unlikely due to the subsequent SCID interview given by a trained psychiatric nurse. Given that this interview was given to all MDD cases in GS:SFHS, misclassification would be less likely between single-episode MDD versus recurrent MDD. Second, publication bias could have influenced results from meta-analyses. Our sample size, although the second largest to investigate MDD and cognition to date, may be underpowered to detect small differences in cognitive performance. Although we removed individuals with Alzheimer and Parkinson's disease and controlled for smoking and alcohol intake, we did not control for other disorders that may affect cognition. Many previous studies focused on clinical populations, whereas our study is population-based; clinical populations may have more severe forms of depression. The use of simpler models in meta-analyses, which do not control for covariates, may obscure signals. Finally, observed cognitive performance in MDD in the literature are mainly observed in large meta-analyses which increases the study heterogeneity, while our results are derived from a much more homogeneous single cohort study. However, both (Snyder, 2013)and (Lim *et al.*, 2013) observed significant heterogeneity and subsequently applied random-effects meta-analytic models that do not assume homogeneity of effect between studies. We also were not able to assess all cognitive domains which could show signs of cognitive impairments in MDD, such as visuospatial processing and attention (Lim *et al.*, 2013). Finally, we were unable to control for the effects of antidepressant use on cognitive performance in the full sample, which may lead to poorer performance in cases.

Cognitive differences between single-episode and recurrent MDD have been not as well studied as differences between MDD cases and controls (Talarowska, Zajaczkowska and Galecki, 2015; Lyall *et al.*, 2016). Talarowska et al. (Talarowska, Zajaczkowska and Galecki, 2015) compared the cognitive performance of 210 patients with MDD (single-episode n=60, recurrent n=150) and observed that the cognitive domains of executive functioning, memory and processing speed showed significant lower performance in recurrent MDD in relation to single-episode MDD.

The largest study to date to assess cognitive differences between single-episode and recurrent depression has been the UK Biobank study (Cullen *et al.*, 2015). Cullen et al, (2015) observed higher performance in single-episode MDD vs controls (numeric and prospective memory), however moderate and severe MDD groups performed lower (e.g. reaction time and numeric memory) compared to both the single MDD and control group.

Cullen et al, 2015 observed the same counter-intuitive higher performance in vocabulary for MDD cases compared to controls and provided several possible explanations for this. It may include differential selection (depressed individuals are more likely to participate than controls), differential recall (cognitive test is associated with greater recall), higher health literacy (individuals with a higher intelligence are quicker to spot possible health issues and therefore quicker to see a GP) or residual confounding.(Cullen *et al.*, 2015) As vocabulary is a crystallised intelligence measure where the tests demand recall ability, and as we observed the same higher performance in a second large population-based cohort, we hypothesise that differential recall and higher health literacy are the most plausible explanations.

That we did not observe a significant genome wide hit for MDD was unsurprising as it is a clinically heterogeneous disorder with multiple SNPs of small effect, which would be difficult to observe without very large sample sizes. We controlled for LD structure in GWAS/GWEIS by applying a less conservative GEC significance threshold which takes into account LD between SNPs. We compared *p*-values of SNPs associated with depression in a large cohort study (Wray and Sullivan, 2018) with our results from the GWEIS studies (Supplementary Table S2.10). Four SNPs overlapped with those available in Generation Scotland and for 18 SNPs we used 52 proxy SNPs ($r^2 > 0.8$). We observed a consistent positive association with *p*-value < 0.05 for the GWEIS of MHVS (both case-control and control-recurrent) and for the GWEIS of DST in control-recurrent analyses for SNP rs4143229 which is intronic and located in *ENOX1*. A recent GWAS of antidepressant treatment response at 12

weeks to selective serotonin reuptake inhibitors (SSRIs) showed suggestive association with another intronic SNP in *ENOX1*, rs17538444. Using Quanto for gene-by-environment power calculations, setting $\alpha = 0.05$, two-sided, and using a MAF of 0.5 (as our top SNP had a MAF of 0.48), and observed MDD proportion and distribution of DST, we concluded that we need a sample size of 2885 individuals was required to detect an interaction effect at 80% power. Although a significant amount of variation in DST was explained by the CHARGE consortium DST polygenic score, it was not specific to MDD cases and the effect did not vary by MDD case status. Polygenic scores often explain only a small amount of variation in endophenotypes. In this study, we looked for main and polygenic effects, it might be possible that more variation can be explained by incorporating possible genetic interactions between loci and/or the environment or interactions of two or more loci.

The main strength of this study is that is has assessed the association between MDD and cognitive ability in a large homogeneous population sample, using standardised tests and outcome measures across all participants. This represents a significant advantage over previous studies that used either meta-analytic (combination of effects across studies) or mega-analytic (combining individual-level data across studies) methods to improve statistical power. The division of the dataset in three study designs based on MDD diagnosis allowed us to assess cognitive performance based on MDD severity. Limitations of this study are the sample size (n=7,012) which results in a low powered interaction analysis, underreporting of antidepressant and mood stabiliser medication usage (<40%) and finally certain cognitive domains are not measured in the Generation Scotland cognitive battery i.e. visuospatial perception.

In conclusion, we have shown that cognitive performance in some domains significantly differs between controls and MDD groups but also within MDD groups. This difference could not be linked to single locus associations but a small proportion of variation could be explained by means of a polygenic approach.

2.6 References

Bora, E. and Pantelis, C. (2015) ‘Meta-analysis of cognitive impairment in first-episode bipolar disorder: Comparison with first-episode schizophrenia and healthy controls’, *Schizophrenia Bulletin*, pp. 1095–1104. doi: 10.1093/schbul/sbu198.

Bowie, C. R. and Harvey, P. D. (2006) ‘Cognitive deficits and functional outcome in schizophrenia’, *Neuropsychiatric Disease and Treatment*, pp. 531–536. doi: 10.2147/nedt.2006.2.4.531.

Clarke, T.-K. *et al.* (2015) ‘Major depressive disorder and current psychological distress moderate the effect of polygenic risk for obesity on body mass index’, *Translational Psychiatry*, 5(6), p. e592. doi: 10.1038/tp.2015.83.

Cullen, B. *et al.* (2015) ‘Cognitive function and lifetime features of depression and bipolar disorder in a large population sample: Cross-sectional study of 143,828 UK Biobank participants’, *European Psychiatry*, 30(8), pp. 950–958. doi: 10.1016/j.eurpsy.2015.08.006.

Delgado, V. B., Kapczinski, F. and Chaves, M. L. F. (2012) ‘Memory mood congruency phenomenon in bipolar I disorder and major depression disorder patients’, *Brazilian Journal of Medical and Biological Research*. Associação Brasileira de Divulgação Científica, 45(9), pp. 856–861. doi: 10.1590/S0100-879X2012007500098.

Fernandez-Pujals, A. M. *et al.* (2015) ‘Epidemiology and heritability of major depressive disorder, stratified by age of onset, sex, and illness course in generation Scotland: Scottish family health study (GS: SFHS)’, *PLoS ONE*. Edited by K. Ebmeier, 10(11), p. e0142197. doi: 10.1371/journal.pone.0142197.

First, M. B. et *et al.* (1997) *Structured Clinical Interview for DSM-IV Axis I Disorders, Clinician Version (SCID-CV), for DSMIV*.

Gooren, T., Schlattmann, P. and Neu, P. (2013) ‘A comparison of cognitive functioning in acute schizophrenia and depression.’, *Acta neuropsychiatrica*, 25(6), pp. 334–41. doi: 10.1017/neu.2013.21.

Halvorsen, M. *et al.* (2012) 'Cognitive function in unipolar major depression: A comparison of currently depressed, previously depressed, and never depressed individuals', *Journal of Clinical and Experimental Neuropsychology*, 34(7), pp. 782–790. doi: 10.1080/13803395.2012.683853.

Ibrahim-Verbaas, C. A. *et al.* (2016) 'GWAS for executive function and processing speed suggests involvement of the CADM2 gene', *Molecular Psychiatry*, 21(2), pp. 189–197. doi: 10.1038/mp.2015.37.

Kendler, K. S. *et al.* (2006) 'A Swedish national twin study of lifetime major depression', *American Journal of Psychiatry*, 163, pp. 109–114. doi: 10.1176/appi.ajp.163.1.109.

Lezak, M.; Howieson, D.; Bigler, E.; Tranel, D. (2012) *Neuropsychological Assessment. 5.* Oxford University Press; 2012).

Li, M. X. *et al.* (2012) 'Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets', *Human Genetics*, 131(5), pp. 747–756. doi: 10.1007/s00439-011-1118-2.

Lim, J. *et al.* (2013) 'Sensitivity of cognitive tests in four cognitive domains in discriminating MDD patients from healthy controls: a meta-analysis', *International Psychogeriatrics*, 25(9), pp. 1543–1557. doi: 10.1017/S1041610213000689.

Lubke, G. H. *et al.* (2012) 'Estimating the genetic variance of major depressive disorder due to all single nucleotide polymorphisms', *Biological Psychiatry*, 72(8), pp. 707–709. doi: 10.1016/j.biopsych.2012.03.011.

Lyall, D. M. *et al.* (2016) 'Cognitive Test Scores in UK Biobank: Data Reduction in 480,416 Participants and Longitudinal Stability in 20,346 Participants', *PLOS ONE*. Edited by H. Reddy. Public Library of Science, 11(4), p. e0154222. doi: 10.1371/journal.pone.0154222.

Maeshima, H. *et al.* (2012) 'Residual memory dysfunction in recurrent major depressive disorder—A longitudinal study from Juntendo University Mood Disorder

Project', *Journal of Affective Disorders*, 143(1–3), pp. 84–88. doi: 10.1016/j.jad.2012.05.033.

Maeshima, H. *et al.* (2013) 'Time course for memory dysfunction in early-life and late-life major depression: A longitudinal study from the Juntendo university mood disorder project', *Journal of Affective Disorders*, 151(1), pp. 66–70. doi: 10.1016/j.jad.2013.05.050.

Payne, R. a and Abel, G. a (2012) 'UK indices of multiple deprivation - a way to make comparisons across constituent countries easier.', *Health statistics quarterly / Office for National Statistics*, (53), pp. 22–37. doi: 10.1017/CBO9781107415324.004.

R Core Team (2017) 'R Core Team (2017). R: A language and environment for statistical computing.', *R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.*, p. R Foundation for Statistical Computing.

Raven JC, Court JH, R. J. (1977) 'Manual for Raven's Progressive Matrices and Vocabulary Scales', *HK Lewis*.

Ripke, S. *et al.* (2013) 'A mega-analysis of genome-wide association studies for major depressive disorder', *Molecular Psychiatry*, 18(4), pp. 497–511. doi: 10.1038/mp.2012.21.

Smith, B. H. *et al.* (2006) 'Generation Scotland: the Scottish Family Health Study; a new resource for researching genes and heritability.', *BMC medical genetics*, 7(1), p. 74. doi: 10.1186/1471-2350-7-74.

Smith, B. H. *et al.* (2013) 'Cohort profile: Generation scotland: Scottish family health study (GS: SFHS). The study, its participants and their potential for genetic research on health and illness', *International Journal of Epidemiology*, 42(3), pp. 689–700. doi: 10.1093/ije/dys084.

Smith, D. J. *et al.* (2013) 'Prevalence and characteristics of probable major depression and bipolar disorder within UK Biobank: Cross-sectional study of 172,751 participants', *PLoS ONE*. Edited by J. B. Potash. Public Library of Science,

8(11), p. e75362. doi: 10.1371/journal.pone.0075362.

Snyder, H. R. (2013) ‘Major depressive disorder is associated with broad impairments on neuropsychological measures of executive function: A meta-analysis and review.’, *Psychological Bulletin*. NIH Public Access, 139(1), pp. 81–132. doi: 10.1037/a0028727.

Talarowska, M., Zajaczkowska, M. and Galecki, P. (2015) ‘Cognitive functions in first-episode depression and recurrent depressive disorder’, *Psychiatria Danubina*, 27(1), pp. 38–43.

Travis, S. *et al.* (2014) ‘Dentate gyrus volume and memory performance in major depressive disorder.’, *Journal of affective disorders*, 172C, pp. 159–164. doi: 10.1016/j.jad.2014.09.048.

Wechsler D. (1998a) ‘WAIS-III UK Wechsler Adult Intelligence Scale.’, *Psychological Corporation*.

Wechsler D. (1998b) ‘WMS-III UK, Wechsler Memory Scale-Revised.’, *Psychological Corporation*.

Wray, N. R. and Sullivan, P. F. (2018) ‘Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression’, *Nature Genetics*. doi: 10.1101/167577.

Yang, J. *et al.* (2011) ‘GCTA: a tool for genome-wide complex trait analysis.’, *American journal of human genetics*, 88(1), pp. 76–82. doi: 10.1016/j.ajhg.2010.11.011.

2.7 Supplementary material

	LM1	DST	VFT	MHVS	LM2
LM1	1				
DST	0.29 (< 2.2E-16)	1			
VFT	0.15 (< 2.2E-16)	0.25 (< 2.2E-16)	1		
MHVS	0.23 (< 2.2E-16)	0.08 (7.01E-10)	0.39 (< 2.2E-16)	1	
LM2	0.86 (< 2.2E-16)	0.30 (< 2.2E-16)	0.15 (< 2.2E-16)	0.21 (< 2.2E-16)	1
LM1-LM2	0.07 (5.91E-08)	-0.08 (2.34E-10)	-0.04 (7.1E-04)	-0.008 (0.5)	-0.44 (< 2.2E-16)

Supplementary Table S2.1: Correlation (*p*-values) between cognitive performance tests for controls.
Correlations between continuous variables were calculated using the Pearson correlation coefficient. LM = logical memory, DST = digit symbol test, VFT = verbal fluency test, MHVS = Mill-Hill Vocabulary test.

	LM1	DST	VFT	MHVS	LM2
LM1	1				
DST	0.34 (< 2.2E-16)	1			
VFT	0.22 (5.5E-13)	0.25 (< 2.2E-16)	1		
MHVS	0.27 (< 2.2E-16)	0.11 (4.1E-04)	0.39 (< 2.2E-16)	1	
LM2	0.86 (< 2.2E-16)	0.35 (< 2.2E-16)	0.22 (1.1E-12)	0.25 (< 2.2E-16)	1
LM1-LM2	0.15 (5.3E-07)	-0.07 (2.4E-02)	-0.02 (0.43)	0.01 (0.7)	-0.35 (< 2.2E-16)

Supplementary Table S2.2: Correlation (*p*-values) between cognitive performance tests for MDD cases.
Correlations between continuous variables were calculated using the Pearson correlation coefficient. LM = logical memory, DST = digit symbol test, VFT = verbal fluency test, MHVS = Mill-Hill Vocabulary test.

	LM1	DST	VFT	MHVS	LM2
LM1	1				
DST	0.31 (1.2E-12)	1			
VFT	0.20(5.8E-06)	0.22(5.4E-07)	1		
MHVS	0.24 (3.6E-08)	0.05 (0.27)	0.39 (< 2.2E-16)	1	
LM2	0.85 (< 2.2E-16)	0.29 (5.3E-11)	0.2 (1.4E-05)	0.24 (1.1E-07)	1
LM1-LM2	0.20 (7.5E-06)	0.01 (0.8)	-0.01 (0.8)	0.02 (0.7)	-0.34 (2.4E-14)

Supplementary Table S2.3: Correlation (*p*-values) between cognitive performance tests for single episode MDD cases.
Correlations between continuous variables were calculated using the Pearson correlation coefficient. LM = logical memory, DST = digit symbol test, VFT = verbal fluency test, MHVS = Mill-Hill Vocabulary test.

	LM1	DST	VFT	MHVS	LM2
LM1	1				
DST	0.35 (2.26E-16)	1			
VFT	0.24 (2E-08)	0.28 (5.4E-11)	1		
MHVS	0.3 (< 3.4E-12)	0.17 (6.7E-05)	0.41 (< 2.2E-16)	1	
LM2	0.87 (< 2.2E-16)	0.4 (< 2.2E-16)	0.24 (1.8E-08)	0.27 (< 1.88E-10)	1
LM1-LM2	0.12 (6E-03)	-0.14 (9.6E-04)	-0.04 (0.41)	0.007 (0.9)	-0.38 (< 2.2E-16)

Supplementary Table S2.4: Correlation (*p*-values) between cognitive performance tests for recurrent episode MDD cases.
Correlations between continuous variables were calculated using the Pearson correlation coefficient. LM = logical memory, DST = digit symbol test, VFT = verbal fluency test, MHVS = Mill-Hill Vocabulary test.

	Control-MDD	Sex	Age	Alcohol	Smoking	Medication
Control-MDD	1					
Sex	0.12 (5.2E-22)	1				
Age	-0.054 (5.0E-6)	-0.06 (3.2E-6)	1			
Alcohol	-0.07 (3.23E-08)	-0.05 (8.44E-05)	-0.05 (1.54E-05)	1		
Smoking	0.11 (2.49E-18)	-0.03 (0.03)	-0.15 (1.04E-37)	-0.04 (1.83E-03)	1	
Medication	0.29 (1.37E-64)	0.05 (3.0E-03)	0.04 (8.0E-03)	-0.06 (1.09E-03)	0.07 (9.42E-05)	1
SES	-0.08 (1.66E-11)	-0.05 (1.68E-05)	0.14 (5.62E-32)	0.08 (6.85E-11)	-0.19 (1.78E-56)	-0.07 (5.52E-06)

Supplementary Table S2.5: Correlation (*p*-values) between covariates for MDD cases and controls.

Correlations between binary variables were calculated using the phi coefficient and corresponding χ^2 or Fisher's exact test p-value. Fisher's exact test was used when cell sizes in the 2 x 2 contingency table with observed values smaller than 5. Correlations between continuous variables were calculated using the Spearman correlation coefficient. SES = socioeconomic status.

	Single-Recurrent	Sex	Age	Alcohol	Smoking	Medication
Single-Recurrent	1					
Sex	0.03 (0.29)	1				
Age	0.04 (0.18)	-0.06 (0.06)	1			
Alcohol	-0.08 (4.80E-306)	0.04 (0.2)	-0.019 (0.54)	1		
Smoking	0.04 (0.19)	-0.01 (0.76)	-0.18 (1.47E-09)	-0.1 (0.002)	1	
Medication	0.21 (3.79E-06)	-0.03 (0.61)	0.15 (8.5E-04)	-0.08 (0.11)	0.04 (0.47)	1
SES	-0.10 (0.001)	-0.06 (0.03)	0.17 (5.01E-08)	0.06 (0.06)	-0.22 (1.54E-12)	-0.04 (0.33)

Supplementary Table S2.6: Correlation (*p*-values) between covariates for single and recurrent MDD.

Correlations between binary variables were calculated using the phi coefficient and corresponding χ^2 or Fisher's exact test p-value. Fisher's exact test was used when cell sizes in the 2 x 2 contingency table with observed values smaller than 5. Correlations between continuous variables were calculated using the Spearman correlation coefficient. SES = socioeconomic status.

	Control-Recurrent	Sex	Age	Alcohol	Smoking	Medication
Control-Recurrent	1					
Sex	-3.44E-04 (0.977)	1				
Age	-0.03 (-0.03)	0.29 (7.92E-130)	1			
Alcohol	-1.94E-3 (0.87)	0.15 (1.47E-35)	0.25 (1.98E-92)	1		
Smoking	6.54E-03 (0.6)	0.23 (1.45-82)	0.08 (2.88E-12)	0.39 (2.18E-234)	1	
Medication	2.83E-04 (0.98)	0.86 (0)	0.3 (3.27E-140)	0.16 (1.95E-37)	0.22	1
SES	-9.3E-04 (0.94)	0.07 (2.39E-09)	-0.08 (3.41E-12)	-0.04 (5.15E-04)	-7.54E-03 (0.55)	-0.44 (3.75E-301)

Supplementary Table S2.7: Correlation (*p*-values) between covariates for recurrent MDD cases and controls.
Correlations between binary variables were calculated using the phi coefficient and corresponding χ^2 or Fisher's exact test p-value. Fisher's exact test was used when cell sizes in the 2 x 2 contingency table with observed values smaller than 5. Correlations between continuous variables were calculated using the Spearman correlation coefficient. SES = socioeconomic status.

	Control-MDD			Single-Recurrent MDD			Control-Recurrent MDD		
	β	*Pr(>|t|)*	*N*	*β*	*Pr(>|t|)*	*N*	*β*	*Pr(>|t|)*	*N*
LM1*	0.40	0.05	3172	-0.28	0.44	442	0.21	0.47	2947
LM2	0.31	0.17	3167	-0.47	0.22	442	-0.02	0.92	2942
LM1-LM2	6.6E-02	0.56	3167	0.19	0.33	442	0.19	0.22	2942
DST	-1.15	0.13	3167	-1.09	0.41	440	-1.84	0.08	2944
VFT	1.07	0.08	3160	2.36	0.04	441	1.99	0.02	2936
MHVS	**0.66**	**2.96E-03**	**3148**	0.99	0.01	442	**1.07**	**6.0E-04**	**2923**

Supplementary Table S2.8: Association between diagnosis label and cognitive performance in both study designs, after controlling for all covariates including medication usage. **LM1: Logical memory immediate, LM2: logical memory delayed, DST: digit symbol substitution test, VFT: verbal fluency total, MHVS: Mill Hill vocabulary score. Bolded results are significant after Bonferroni correction.*

	Single-Recurrent MDD		Control-Recurrent MDD	
Range	*β*	*Pr(>|t|)*	*β*	*Pr(>|t|)*
0-0.01	$2.36E^{4}$	$9.75E^{-02}$	$1.10E^{-4}$	0.29
0-0.05	$3.15E^{4}$	0.37	$3.08E^{4}$	0.24
0-0.1	$5.42E^{4}$	0.31	$4.95E^{4}$	0.22
0-0.5	$2.31E^{5}$	0.15	$1.06E^{5}$	0.39
0-1	$4.72E^{5}$	0.11	$2.21E^{5}$	0.34

Supplementary Table S2.9. Association between DST performance and PRS*MDD status derived from the DST meta-analysis of the CHARGE consortium.

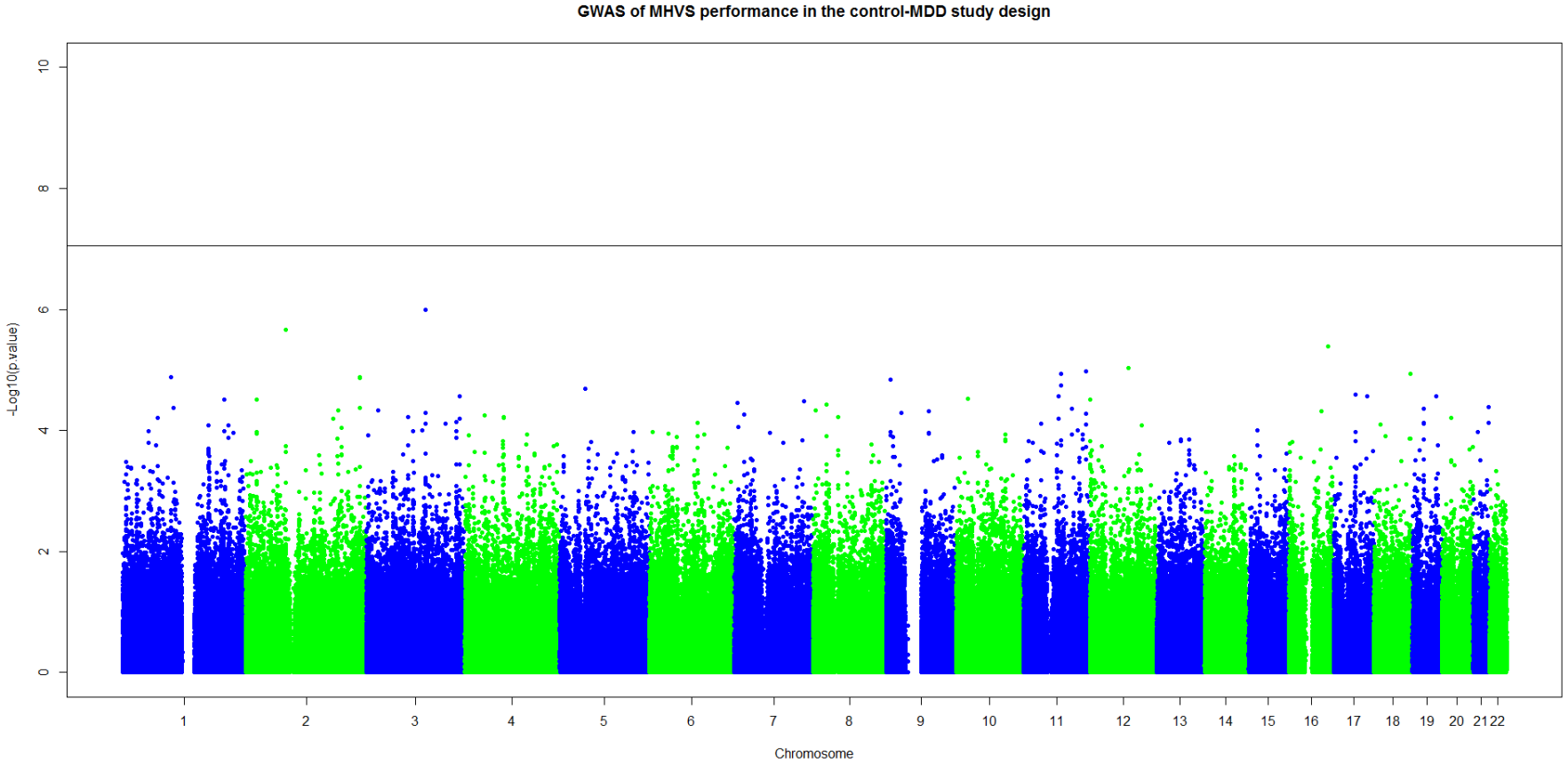
				GWEIS MHVS C-C	**GWEIS MHVS C-R**	**GWEIS DST C-R**	**GWEIS DST S-R**
SNP Wray et al, 2018	*p*	*SNP Generation Scotland*	*r*	*p*	*p*	*p*	*p*
rs4143229	**2.5E-08**	**rs4143229**	**N.A**	**0.0307**	**0.02076**	**0.005283**	**0.176**
rs12552	**6.1E-19**	**rs12552**	**N.A**	**0.521**	**0.853**	**0.6496**	**0.9415**
rs11643192	**3.4E-08**	**rs11643192**	**N.A**	**0.2208**	**0.09025**	**0.4155**	**0.09073**
rs1833288	**2.6E-08**	**rs1833288**	**N.A**	**0.2129**	**0.1569**	**0.32**	**0.9892**
rs159963	3.2E-08	rs301806	1	0.1536	0.8932	0.4014	0.8966
rs159963	3.2E-08	rs301805	0.966	0.1395	0.8448	0.3632	0.8941
rs159963	3.2E-08	rs4908760	0.811	0.3446	0.8495	0.4729	0.4594
rs1432639	4.6E-15	rs2012697	1	0.1228	0.1945	0.2898	0.5964
rs1432639	4.6E-15	rs2568958	1	0.1259	0.1626	0.3438	0.7348
rs1432639	4.6E-15	rs3101336	1	0.1473	0.1567	0.3578	0.7346
rs1432639	4.6E-15	rs2815752	0.962	0.1277	0.1642	0.3576	0.7346
rs1432639	4.6E-15	rs7531118	0.8	0.1125	0.1877	0.4689	0.9
rs12129573	4.0E-12	rs1160682	0.965	0.6188	0.9106	0.1785	0.07364
rs12129573	4.0E-12	rs11210201	0.932	0.7437	0.8731	0.6114	0.4146
rs12129573	4.0E-12	rs1885246	0.932	0.7527	0.8652	0.6126	0.4004
rs12129573	4.0E-12	rs1475064	0.87	0.8984	0.6614	0.4694	0.4062
rs12129573	4.0E-12	rs9425120	0.864	0.9921	0.8598	0.8455	0.6465
rs2389016	1.0E-08	rs12065553	1	0.3443	0.3818	0.8636	0.3692
rs2389016	1.0E-08	rs2389024	1	0.3371	0.3993	0.9061	0.3692
rs2389016	1.0E-08	rs2154298	1	0.35	0.4313	0.9571	0.3262
rs2389016	1.0E-08	rs12118987	1	0.4025	0.5497	0.9244	0.3312
rs2389016	1.0E-08	rs10158964	1	0.3493	0.4594	0.839	0.2491
rs2389016	1.0E-08	rs1937787	1	0.434	0.5931	0.8752	0.2543
rs2389016	1.0E-08	rs3856038	0.965	0.4188	0.6911	0.9654	0.2772

rs2389016	1.0E-08	rs10493662	0.964	0.3594	0.3923	0.8493	0.3774
rs4261101	1.0E-08	rs4453022	0.963	0.2519	0.5699	0.3363	0.215
rs12958048	3.6E-11	rs4468713	0.858	0.569	0.7404	0.05335	0.2768
rs4074723	3.1E-08	rs4511370	0.869	0.6151	0.7961	0.84	0.9681
rs7430565	2.9E-09	rs6774461	1	0.1329	0.7494	0.1807	0.105
rs7430565	2.9E-09	rs2682405	0.967	0.1807	0.812	0.2691	0.1636
rs7430565	2.9E-09	rs7643792	0.967	0.1663	0.8058	0.2697	0.184
rs7430565	2.9E-09	rs1213048	0.967	0.1704	0.8108	0.2706	0.1827
rs7430565	2.9E-09	rs9857883	0.967	0.2292	0.8517	0.2798	0.2048
rs11135349	1.1E-09	rs10866752	0.81	0.3304	0.3124	0.3151	0.6401
rs4869056	6.8E-09	rs11747772	0.931	0.3579	0.3431	0.5339	0.4227
rs4869056	6.8E-09	rs11738110	0.899	0.3959	0.3636	0.6622	0.5927
rs4869056	6.8E-09	rs883322	0.834	0.4268	0.2816	0.758	0.5952
rs12666117	1.4E-08	rs11561993	0.967	0.9872	0.2198	0.1912	0.06189
rs12666117	1.4E-08	rs12113865	0.837	0.7409	0.2728	0.4964	0.09299
rs1354115	4.7E-09	rs7044150	0.965	0.3715	0.4744	0.3247	0.1761
rs1354115	4.7E-09	rs4741790	0.965	0.2827	0.4369	0.4942	0.4155
rs1354115	4.7E-09	rs7033160	0.868	0.8874	0.606	0.5845	0.241
rs1354115	4.7E-09	rs4741798	0.801	0.8595	0.7233	0.4806	0.2342
rs7856424	8.5E-09	rs10759879	1	0.9186	0.3242	0.8277	0.6453
rs61867293	7.0E-10	rs11192270	0.865	0.7957	0.1699	0.303	0.7227
rs61867293	7.0E-10	rs10884071	0.826	0.8955	0.5312	0.5209	0.7894
rs61867293	7.0E-10	rs17766570	0.802	0.424	0.1119	0.4206	0.5602
rs4904738	2.6E-09	rs1983711	0.93	0.4879	0.1866	0.2595	0.9982
rs915057	7.6E-10	rs7229	0.966	0.7609	0.3643	0.7707	0.9363
rs8025231	2.4E-12	rs668644	0.846	0.6272	0.4287	0.3206	0.9017
rs8025231	2.4E-12	rs624991	0.818	0.6517	0.4061	0.306	0.8907

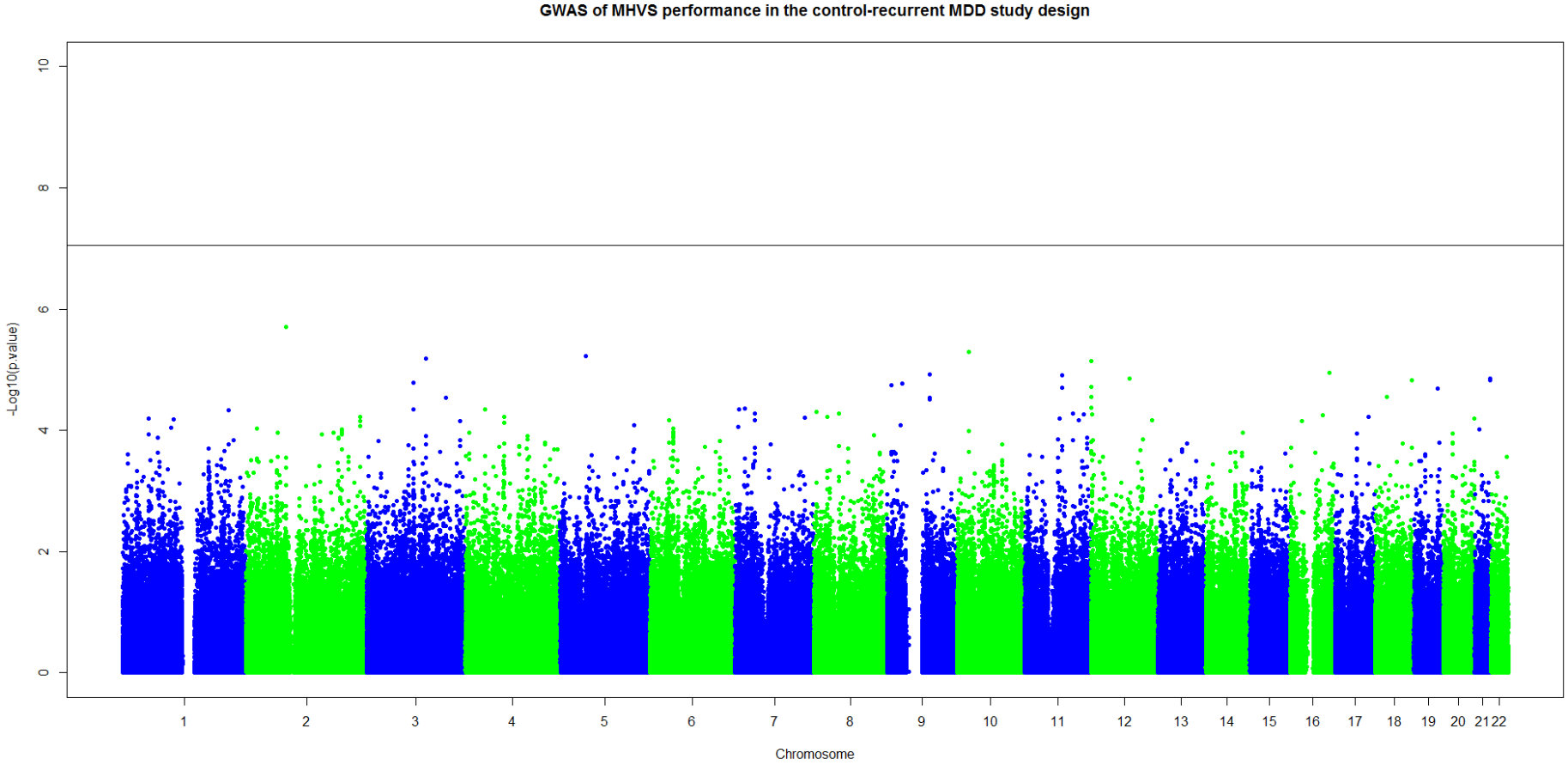
rs8025231	2.4E-12	rs8024814	0.815	0.6912	0.6527	0.3112	0.9395
rs8025231	2.4E-12	rs657586	0.815	0.7187	0.636	0.298	0.9228
rs8025231	2.4E-12	rs667471	0.815	0.7493	0.7686	0.29	0.9647
rs7198928	1.0E-08	rs11077203	1	0.2905	0.2424	0.1135	0.1331
rs7198928	1.0E-08	rs7192025	0.963	0.2471	0.2457	0.1627	0.09317

Supplementary Table S2.10: Comparison between Wray et al, 2017 observed results and Generation Scotland equivalent SNPs.
r = Correlation between SNP Wray et al, 2017 and Generation Scotland SNP, bolded SNPs overlap between Wray et al, 2017 and Generation Scotland.

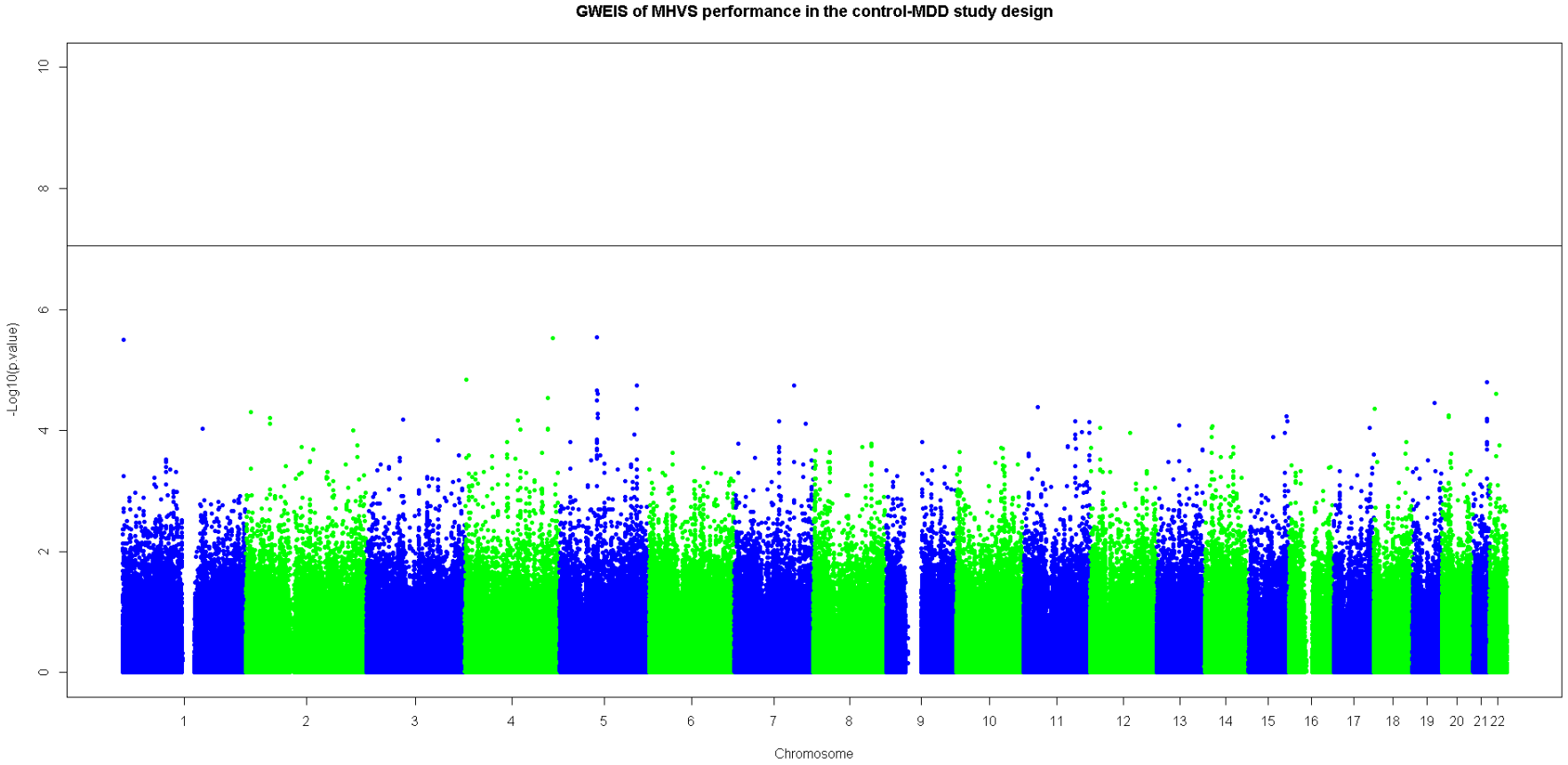
For all subsequent GWAS/GWEIS figures: horizontal black line indicates Bonferroni threshold of $p = 5\times10^{-08}$



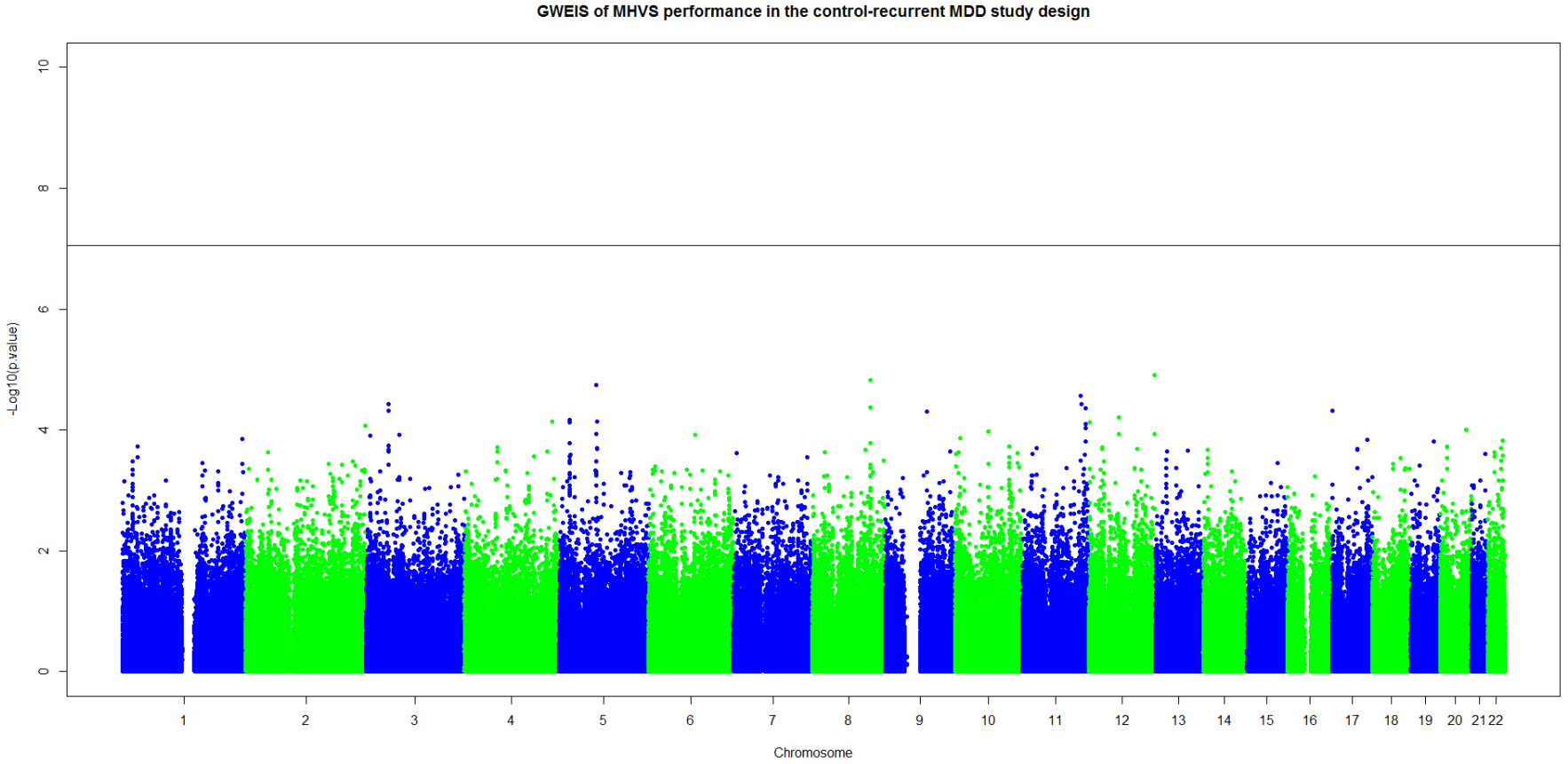
Supplementary Figure S2.1A: GWAS of MHVS in the control-MDD study design controlling for all covariates except medication usage



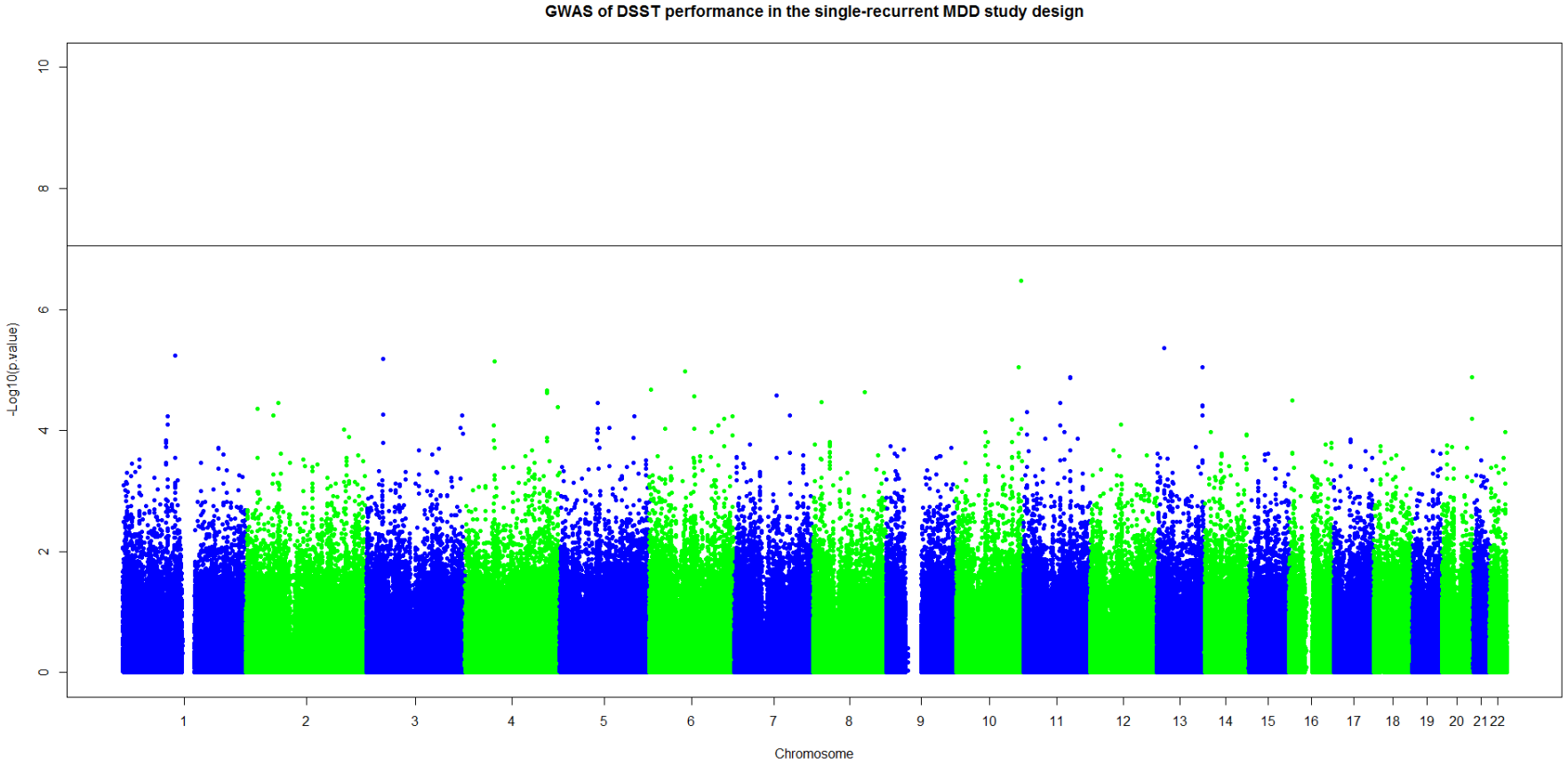
Supplementary Figure S2.1B: GWAS of MHVS in the control-recurrent MDD study design controlling for all covariates except medication usage



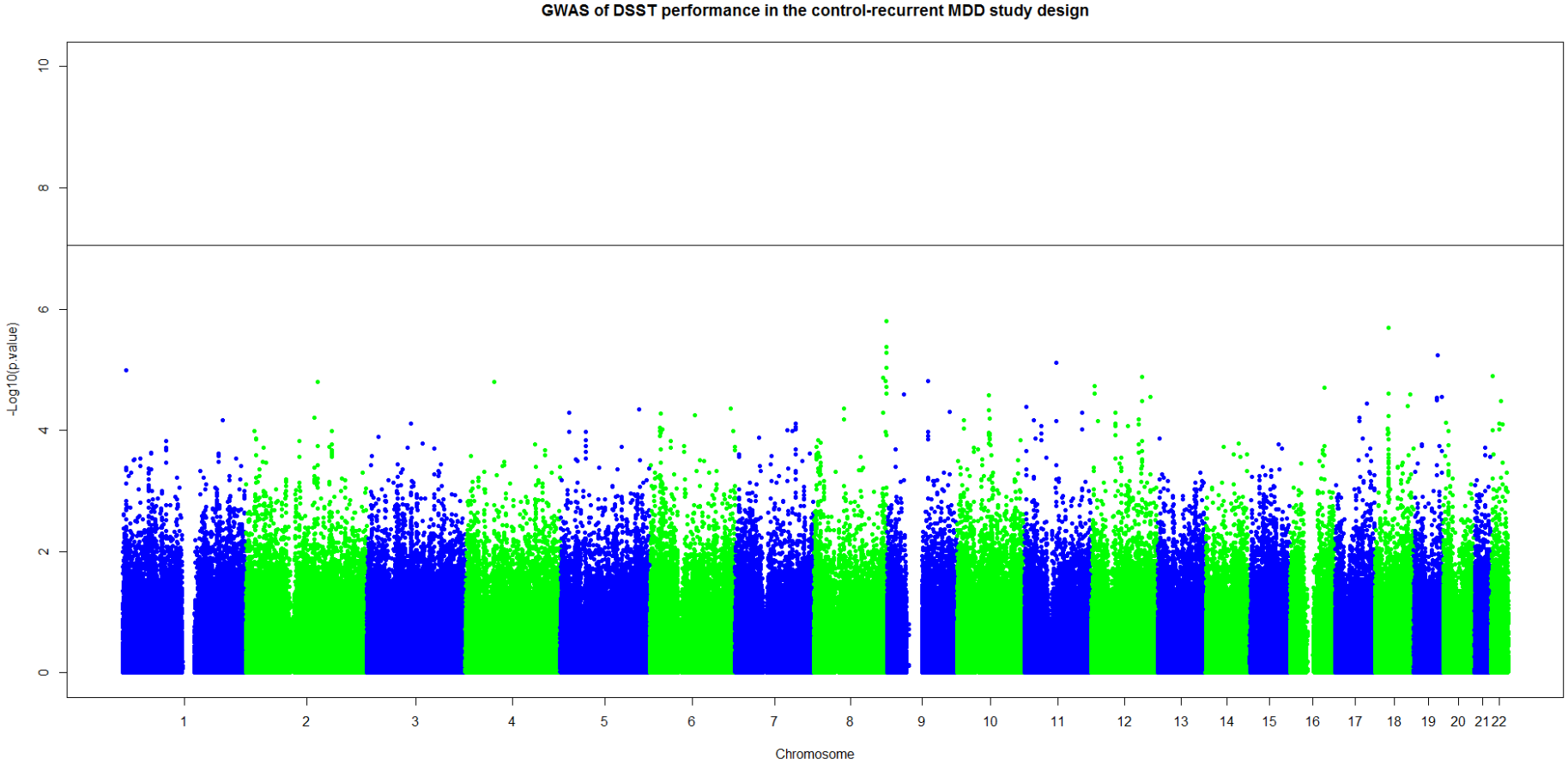
Supplementary Figure S2.2A: GWEIS of MHVS in the control-MDD study design controlling for all covariates except medication usage



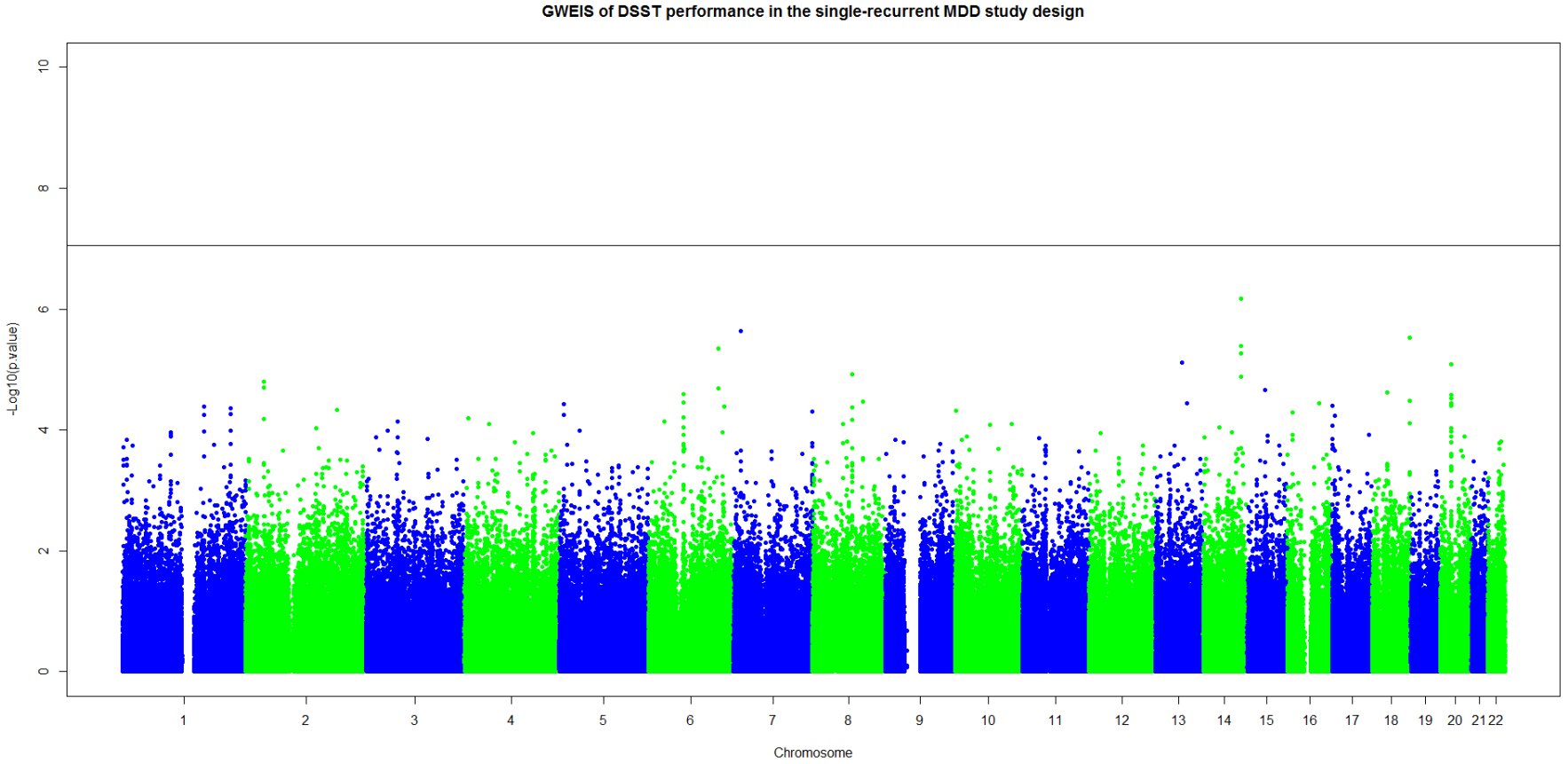
Supplementary Figure S2.2B: GWEIS of MHVS in the control-recurrent MDD study design controlling for all covariates except medication usage



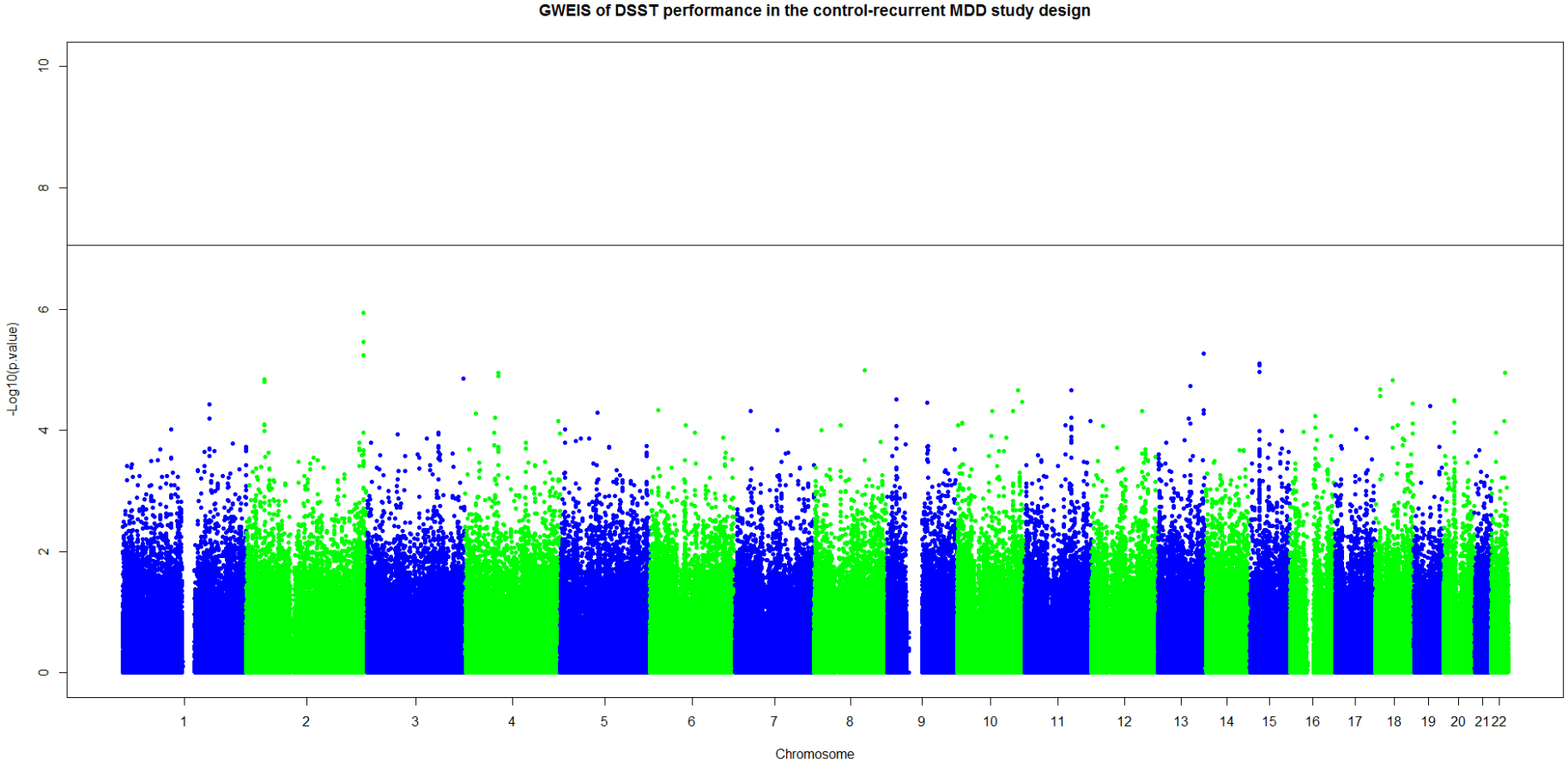
Supplementary Figure S2.3A: GWAS of DST in the single-recurrent MDD study design controlling for all covariates except medication usage



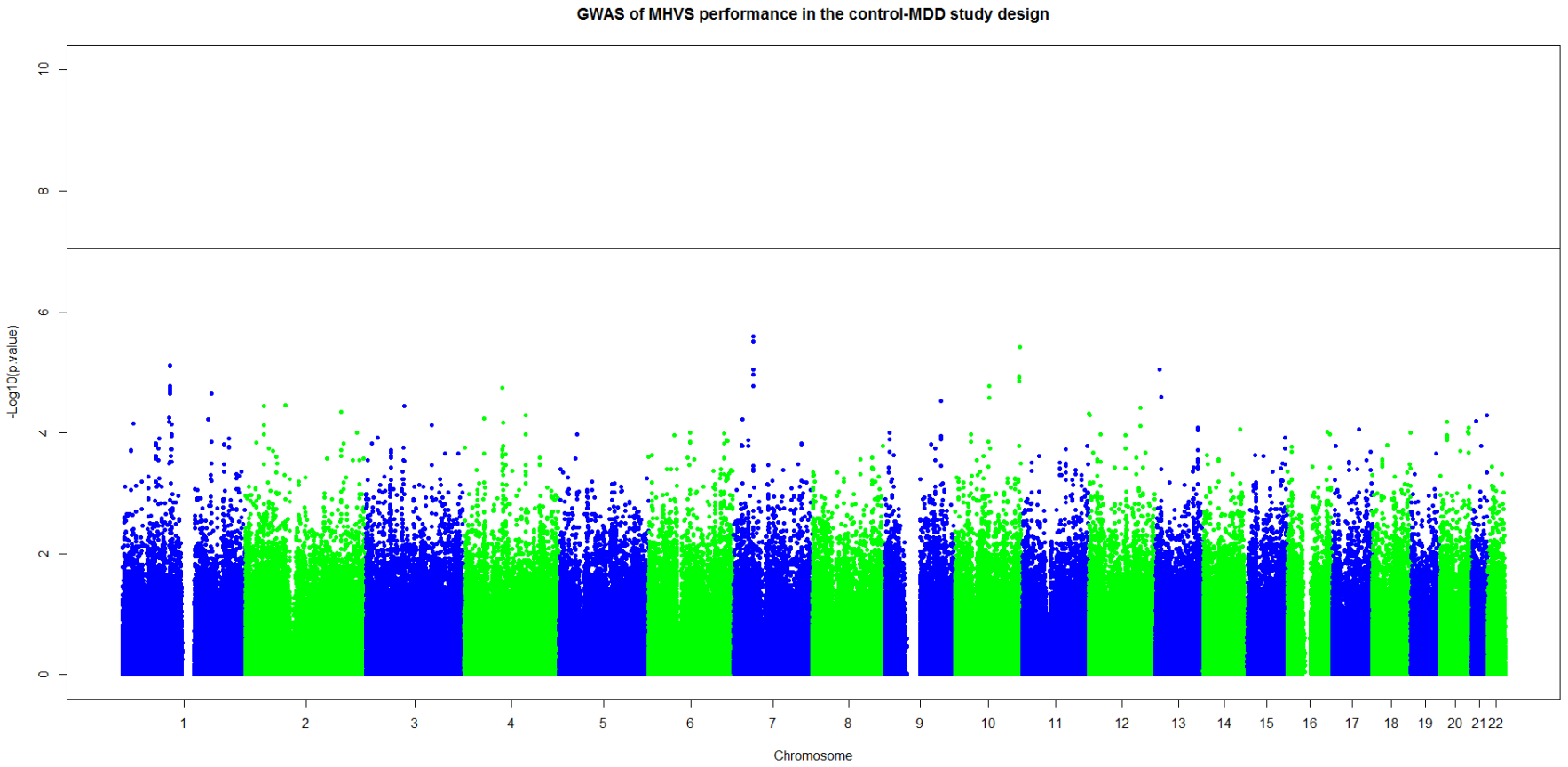
Supplementary Figure S2.3B: GWAS of DST in the control-recurrent MDD study design controlling for all covariates except medication usage



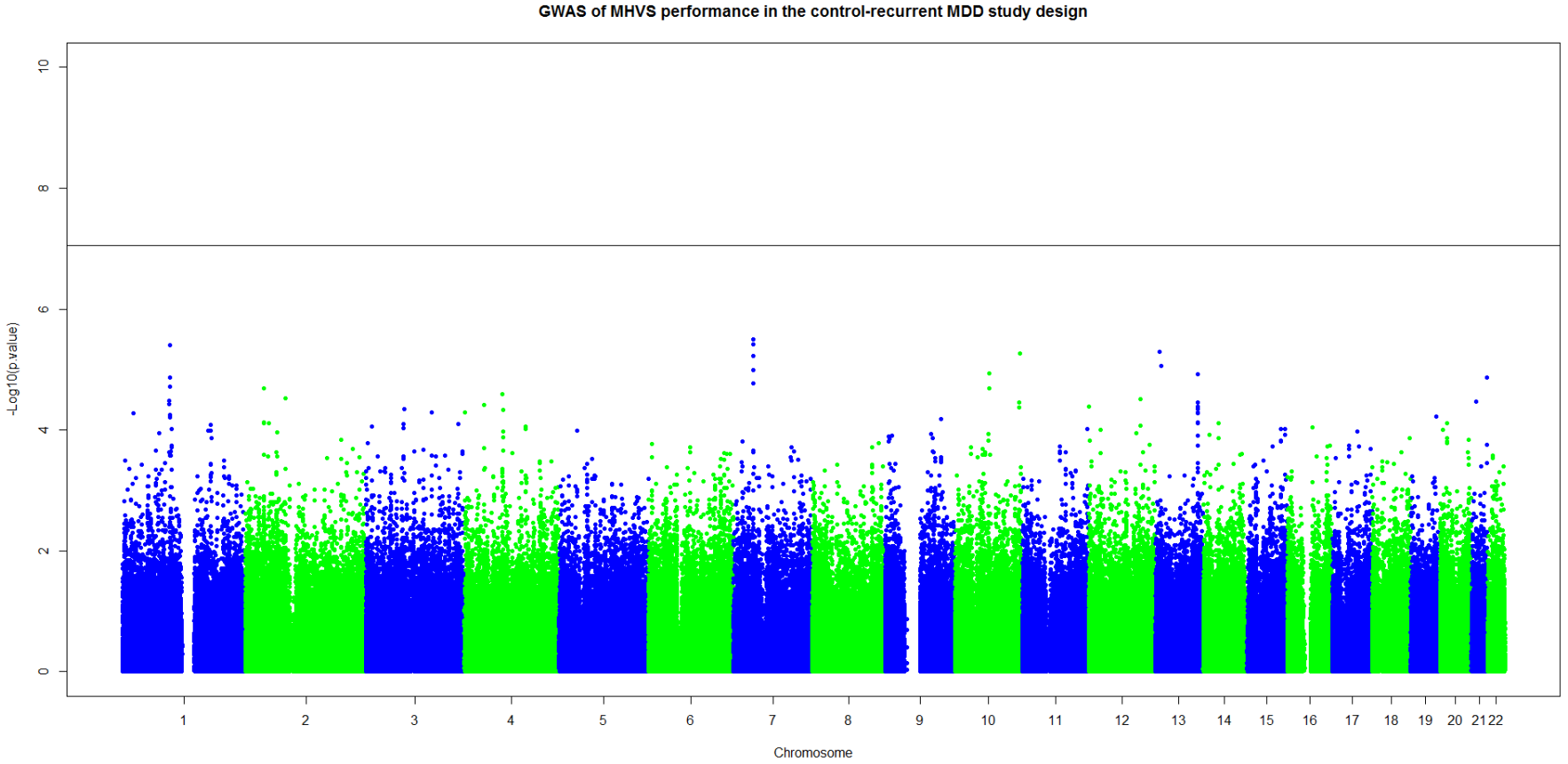
Supplementary Figure S2.4A: GWEIS of DST in the single-recurrent MDD study design controlling for all covariates except medication usage



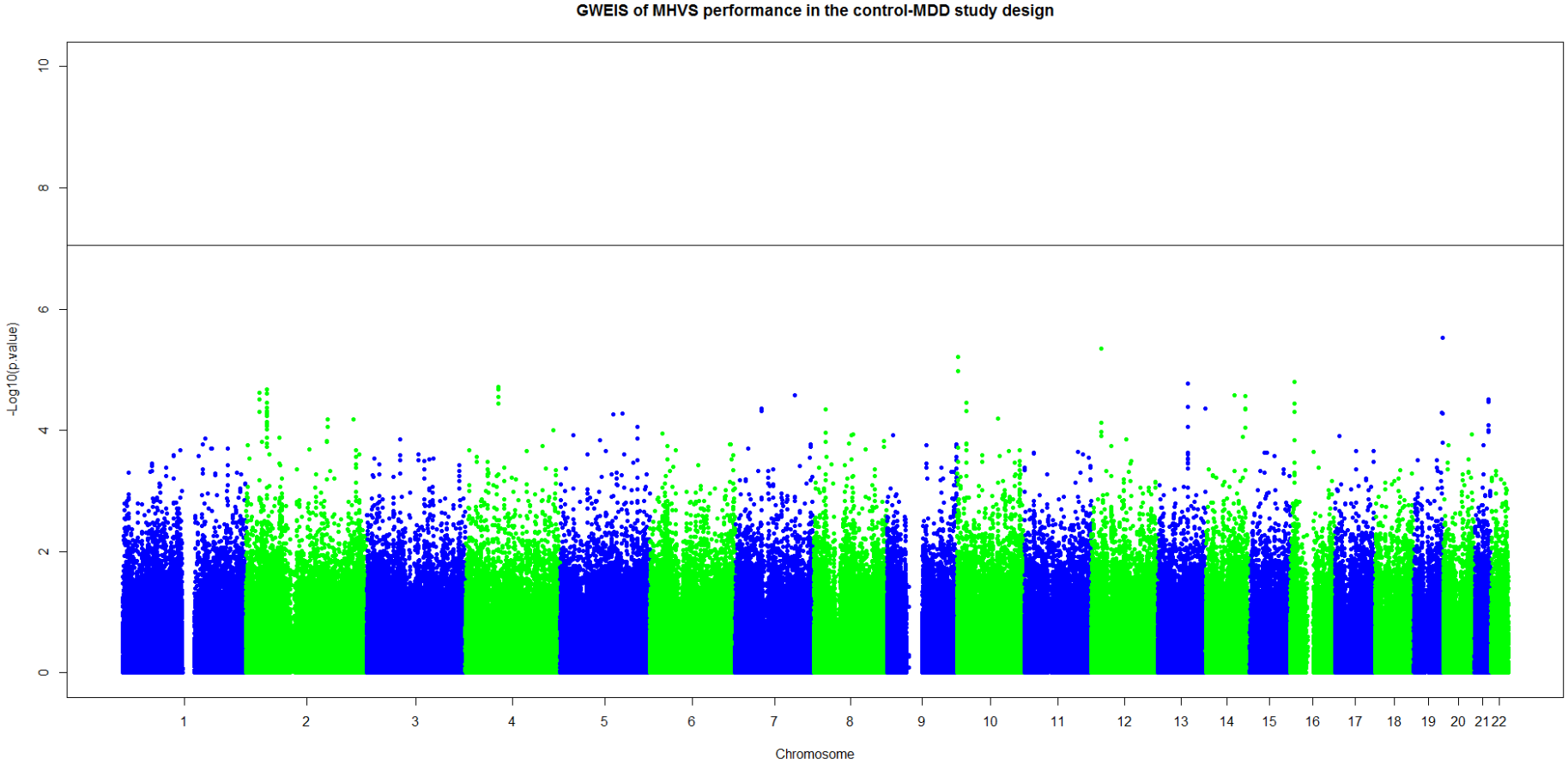
Supplementary Figure S2.4B: GWEIS of DST in the control-recurrent MDD study design controlling for all covariates except medication usage



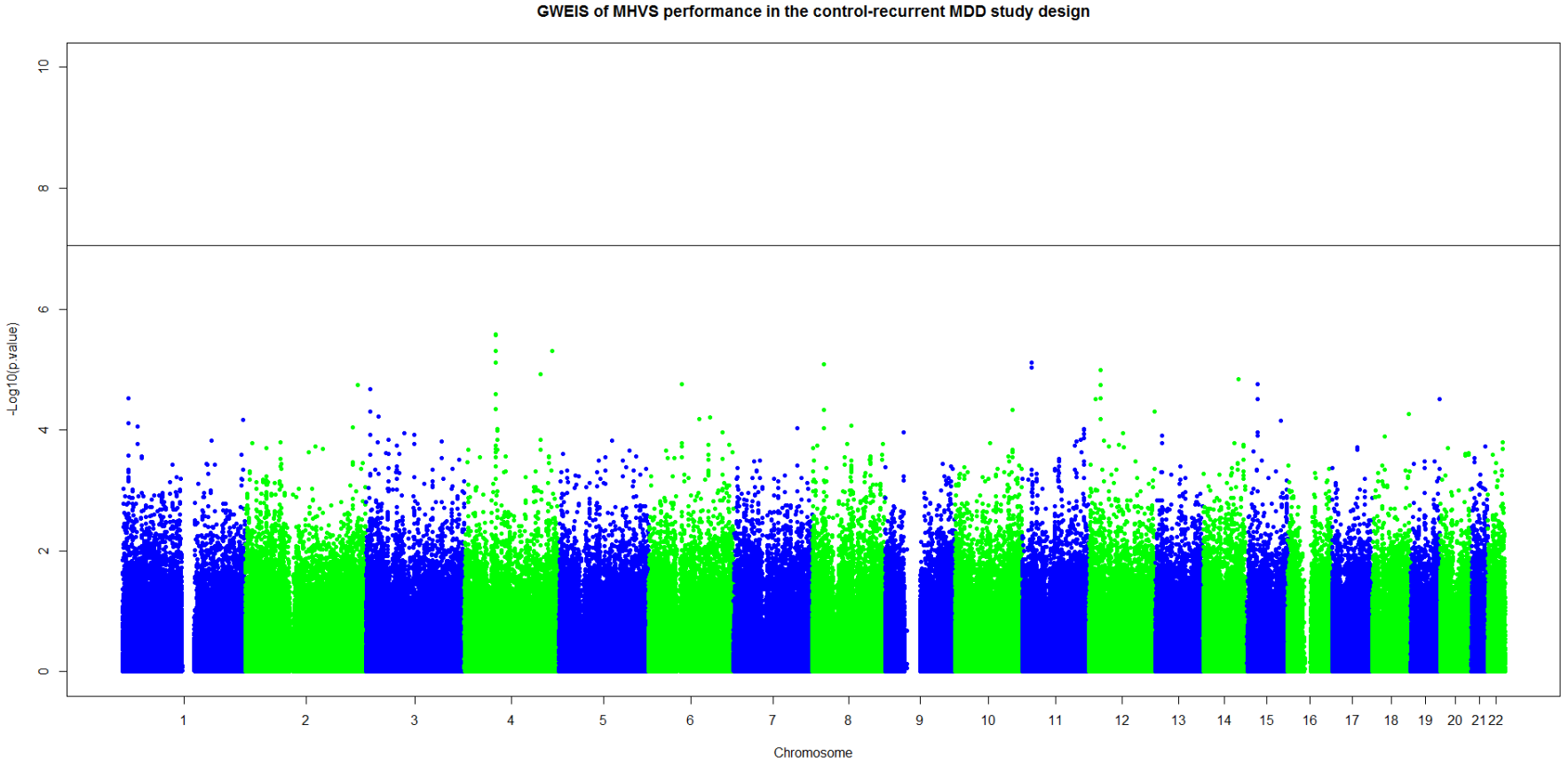
Supplementary Figure S2.5A: GWAS of MHVS in the control-MDD study design controlling for all covariates



Supplementary Figure S2.5B: GWAS of MHVS in the control-recurrent MDD study design controlling for all covariates



Supplementary Figure S2.6A: GWEIS of MHVS in the control-MDD study design controlling for all covariates



Supplementary Figure S2.6b: GWEIS of MHVS in the control-recurrent MDD study design controlling for all covariates

Brief conclusion

This published worked outlined that indeed cognitive performance in MDD cases is significantly different compared to controls, but also between single-episode and recurrent MDD. This difference however depends on the cognitive domain as processing speeds shows a negative difference while vocabulary shows a positive difference. The result of a positive difference in the vocabulary domain between MDD cases and controls was previously also observed in the large UK Biobank study. Observed phenotypic differences between MDD cases and controls could not be contributed to genetic variants using GWAS and GWEIS. Using a summary statistics of large processing speed meta-analysis 1% of variation in processing speeds performance was explained using a polygenic score.

Chapter 3

Using tree-based methods for detection of gene-gene interactions in the presence of a polygenic signal: simulation study with application to educational attainment in the Generation Scotland Cohort Study

Meijsen JJ, Rammos A, Campbell A, Hayward C, Porteous DJ, Deary IJ, Marioni RE, Nicodemus KK.

Work presented in this chapter has been peer reviewed and published in

Bioinformatics 2018 vol: 33 (17) pages: 63. Pages: 2699-2705

doi: 10.1093/bioinformatics/bty462

Brief introduction

This chapter outlines published work on an extensive simulation study of modelling and detecting epistasis in the human genome using two novel non-parametric tree based methods. By using these methods no limits (either statistically or computationally) are set on the number of interactions investigated. Both methods have been applied on years of education as a proxy for education attainment.

Statement outlining the contribution of first author and co-authors

Alex Rammos contributed to this paper by solely simulating the polygenic phenotype. All subsequent simulated data generation, analyses and writing of/for this paper have been done by Joeri Meijsen.

3 Using tree-based methods for detection of gene-gene interactions in the presence of a polygenic signal: simulation study with application to educational attainment in the Generation Scotland Cohort Study

3.1 Abstract

Motivation: The genomic architecture of human complex diseases is thought to be attributable to single markers, polygenic components and epistatic components. No study has examined the ability of tree-based methods to detect epistasis in the presence of a polygenic signal. We sought to apply decision tree-based methods, C5.0 and logic regression, to detect epistasis under several simulated conditions, varying strength of interaction and linkage disequilibrium (LD) structure. We then applied the same methods to the phenotype of educational attainment in a large population cohort. **Results:** LD pruning improved the power and reduced the type I error. C5.0 had a conservative type I error rate whereas logic regression had a type I error rate that exceeded 5%. Despite the more conservative type I error, C5.0 was observed to have higher power than logic regression across several conditions. In the presence of a polygenic signal, power was generally reduced. Applying both methods on educational attainment in a large population cohort yielded numerous interacting SNPs; notably a SNP in *RCAN3* which is associated with reading and spelling and a SNP in *NPAS3*, a neurodevelopmental gene.

3.2 Introduction

Historically, genomic association studies have focused almost exclusively on single-loci and/or polygenic risk score (PGRS) associations. These methods have been very successful; however, frequently they do not explain the total genetic variance of a trait estimated by twin studies. Therefore, it is also important to consider non-additive genetic effects such as epistasis in the complex genetic architecture of human traits. Epistasis has been described as one genetic locus masking or modifying alleles

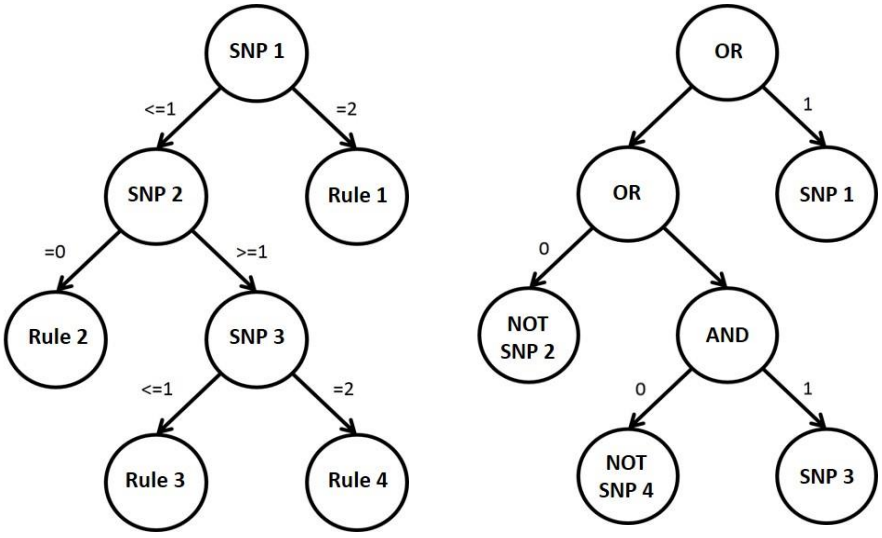
of other loci (Bateson and Mendel, 1909) or a deviation from additivity of two genetic variants on a phenotypic trait (Fisher, 1919). Epistasis, in the sense of 'deviation from additivity' can be defined as either antagonistic (a model where the interaction decreases or blocks the effect of each individual allele) or synergistic (where a combination of alleles exacerbates the effect of each allele individually). Many—if not most—complex traits might have different components of genomic architecture of varying importance—e.g. strongly associated single SNPs, a polygenic component and an epistatic component. The evaluation of statistical learning methodologies for the detection of these different components, to our knowledge, has not been performed.

Even though epistasis has been observed and is well documented in multiple non-human organisms (Brockmann *et al.*, 2000; Cheng *et al.*, 2011; Huang *et al.*, 2012; Mackay, 2014; Grice, Liu and Webber, 2015; He *et al.*, 2016), whether or not epistasis exists and plays a vital role in human traits remains an open debate (Hill, Goddard and Visscher, 2008; Mackay and Moore, 2014; Huang and Mackay, 2016; Sackton and Hartl, 2016; Webber, 2017). According to the 'omnigenic' model, in complex traits the disease-related genetic signal tends to be spread across the genome, resulting in genes without direct statistical association to the trait. Therefore, the 'omnigenic' model states that, due to a large interconnection between gene regulatory networks, most heritability can be explained by the surrounding genes outside the core disease-related genes, which likely includes epistasis (Boyle *et al*., 2017). In general, human epistatic studies have shown limited success, partially due to the use of restrictive methods such as searching within subsets of loci or for specific SNP interaction sizes (e.g. hypothesis-driven analysis) in order to lower the number of tests that need to be performed and thus the resulting statistical correction that has to be applied.

Recently, increasing efforts have been placed on addressing the statistical and computational problems related to the detection of epistasis in large datasets. Machine learning (ML) algorithms are increasingly used to ascertain classifiers for either data reduction or feature selection. These include tree-based methods like random forest (RF), classification and regression trees (CART) (García-Magariños *et al.*, 2009; Chen *et al.*, 2011; Stephan, Stegle and Beyer, 2015) and likelihood ratio Mann-Whitney tests (Lu *et al.*, 2012). García-Magariños *et al. (*2009) simulated genotype data containing interacting SNPs under multiple scenarios (sample size, missing data, minor allele

frequencies and several penetrance models). This study found that CART and RF were equally good in detecting interacting SNPs. Even though the study simulated 99 different scenarios with 100 replicates each, the simulated datasets are very small (two 'causal' SNPs plus 98 null SNPs) and do not reflect the scale or complexity of modern genomic studies.

We sought to apply greedy non-parametric decision tree-based methods—C5.0 and logic regression—for the detection of epistasis in large-scale studies, as these methods explicitly model interactions. C5.0 constructs rule-based decision trees using solely the Boolean operator OR (Figure 3.1A) whereas logic regression allows for Boolean operators AND, OR and NOT (Figure 3.1B). Note that logic regression is a regression framework therefore allowing for the construction of multiple trees (e.g. multiple trees acting as predictors in a regression model), whereas C5.0 constructs multiple rulesets and is not embedded in a regression framework. To date, C5.0 has never been applied to genetic data in the search for interactions, whereas logic regression has been shown to be effective in detecting main effects and interactions in genetic data and could be used as a comparison method (Kooperberg *et al.*, 2001; Ruczinski, Kooperberg and LeBlanc, 2003, 2004; Schwender and Ickstadt, 2008; Chen *et al.*, 2011).



A. **B.**

Figure 3.1: Visual representation of a C5.0 and logic tree. (A) C5.0 decision tree; (B) logic tree

We sought a complex, but well studied trait to test these approaches. Educational attainment (EA) is a highly heritable complex trait (Calvin *et al.*, 2012; Krapohl *et al.*, 2014) and is highly influenced by social and other environmental factors; however, SNP-based heritability estimates that genetic factors contribute to around 20% of variation across individuals, while average twin-based heritability is around 40% (Rietveld *et al.*, 2013). The largest GWAS to date investigating years of education as a proxy of EA observed 74 statistically significant SNPs (Okbay *et al.*, 2016) of which 72 were replicated in the same study using the large UK Biobank cohort. PGRS derived from the same GWAS explained 3.9% of variance in years of education in an independent sample. This large gap of missing heritability (Δh^2_{twin}-h^2_{SNP}) is in similar to that found in other complex traits, however the moderate correlation with traits showing evidence of epistatic contribution e.g. personality traits (Jang, Livesley and Vemon, 1996; Loehlin, Neiderhiser and Reiss, 2003; de Moor *et al.*, 2012; Power and Pluess, 2015; Vukasović and Bratko, 2015) hints towards an epistatic contribution.

In this study, we applied C5.0 and logic regression on simulated epistatic data under multiple scenarios to show their capability of detecting interacting loci in a large genetic study. We sought to assess the performance of C5.0 and logic regression to detect epistatic components alone, plus in the presence of a polygenic signal in order to inform about the methodological development of models that include effects of single SNPs, additive or polygenic components as well as epistasis. To our knowledge, this will be the first simulation study to date to examine the detection of epistasis in the presence of a strong polygenic signal. We applied both methods on the genome-wide SNP data from the Generation Scotland: the Scottish Family health Study (GS:SFHS) cohort to investigate whether there is evidence for an epistatic contribution to years of education as a measurement of educational attainment.

3.3 Materials and Methods

3.3.1 Statistical Methodology

3.3.1.1 Classification and Regression Trees

CARTs are decision tree-based methods that can be interpreted as a set of decisions leading along a path to a final prediction. CART methods utilize classifiers (measurements) to 'split' the data into partitions. CART methods solely use the Boolean operator OR to split a classifier (e.g. male OR female). CART methods grow a tree by including classifiers (recursive partitioning), calculating for every split the 'impurity' or misclassification rate, and define a split with the lowest impurity. Commonly-used impurity measurements are the Gini index for classification-based methods and sum of squared residuals for regression-based methods. CART methods keep recursively partitioning the dataset until no split that decreases impurity can be made or when the size of the terminal nodes (e.g. subjects in node) is less than some user-defined value or is 1. This most often leads to a large tree where some terminal nodes only contain a small number of individuals. The complexity of a tree can be decreased by pruning sections of the tree that provide little power to classify observations.

3.3.1.2 *C5.0 and logic regression*

C5.0 is a modified version of Quinlan's non-parametric C4.5 classification algorithm (Quinlan, 1992). C5.0 builds decision trees, performs rule-based models and evaluation of variable importance (Wu *et al.*, 2008; Kuhn and Johnson, 2013). C5.0 decision trees are built by using information entropy (Equation 3.1).

$$info_{before}^{S} = -\sum_{i=1}^{m} p_i \log p_i \qquad (3.1)$$

Where p_i is the probability of a given class i as the outcome for each of m possible classes and S is the split.To build a tree containing optimal splits, C5.0 assesses, for each node, the normalised information gain which acts as the purity criterion. For each

node C5.0 calculates the information entropy before (Equation 3.1) and after (Equation 3.2) a split.

$$info_{after}^{S} = \sum_{i=1}^{k} info_i \frac{n_i}{n} \quad (3.2)$$

Where *S* is the split; *K* is the number of partitions; n_i is the number of samples *i* assigned to partition *K*; *n* is the total number of samples and $info_i$ is the the sum of the information entropy in the *i*th resulting partition.

For a given node with split *S* and *K* partitions, C5.0 calculates the information entropy for each resulting partition. This is subsequently multiplied by the proportion of samples assigned to that partition $\left(\frac{n_i}{n}\right)$. This adds a weight to each partition, which is summed over all partitions resulting in the information entropy after split S. A lower information entropy after the split implies an information gain (positive difference) and therefore a decrease in uncertainty. If entropy increases (negative difference) C5.0 stops adding splits. The information gain is normalized to allow for the consideration of each class. C5.0 then selects the class with the highest normalized information gain. This process is repeated recursively for smaller subsets.

Each top-to-bottom path in the final tree is collapsed into a so called *rule*. C5.0 evaluates each rule on independent conditional statements, thereby assessing whether or not they can be generalized by removing terms in the conditional statement. This process is called rule-based pessimistic pruning and in short removes branches that are not contributing to the improvement of the trees classification. As a final step, C5.0 assigns each rule to a class by calling a vote. The class with the highest vote is used. Results in a single pruned tree where each possible combination from the top node to bottom node in the tree is a so-called ruleset.

Logic regression is a non-parametric adaptive regression method (Ruczinski, Kooperberg and LeBlanc, 2003). Logic regression is largely based on the same principles as a CART, but in contrast to CART, logic regression constructs logic trees (*L*). Logic trees are Boolean combinations (AND, OR and NOT) of binary predictors e.g., L_1 = SNP_3 or [SNP_1 and (not SNP_4 and not SNP_2)]. This increases the complexity compared to CART which solely applies the Boolean operator OR. A logic tree can be used as a predictor in a regression model (Equation 3.3). Due to its adaptive nature,

logic regression estimates the coefficients (βs) and Boolean expressions (Ls) at the same time.

$$Y = \beta_0 + \beta_1 L_1 + \beta_2 L_2 + \ldots\ldots \beta_p L_p \quad (3.3)$$

By doing so, logic regression tries to minimize the scoring function associated with a model type (e.g. residual sum of squares for quantitative outcomes). For the construction of logic trees, logic regression starts at a random starting point and applies a greedy hill climbing algorithm, which keeps adding predictors to the model as long as the misclassification rate goes down and only stops when the misclassification rate goes up.

3.4 Simulation and genetic methodology

3.4.1 Generation Scotland

Generation Scotland: the Scottish Family Health Study (GS:SFHS) is a large, family-based cohort study sampled from the general population in Scotland (www.generationscotland.org). The study design has been widely documented (Smith *et al.*, 2006, 2013). In short, 24 000 individuals were recruited in the study during a five-year period (2006–2011). The individuals were deeply phenotyped for a wide variety of traits such as lifestyle factors, family history and health outcomes. DNA of 20 128 GS: SFHS individuals were analyzed by means of high density genome wide bead array genotyping (Illumina OmniExpress 700K SNP GWAS and 250K exome chip). DNA results of 134 individuals were excluded during quality control leaving 19 994 genotyped individuals.

We removed single nucleotide polymorphisms (SNPs) and individuals >5% missing data and removed SNPs with a minor allele frequency < 1%. We used Genome-wide Complex Trait Analysis (GCTA) (Yang *et al.*, 2011) to extract a list of genetically-unrelated individuals, giving a total of 7372 individuals (relatedness < 0.025, corresponding to second degree cousins). For the simulation study, we selected 5000 individuals at random from the unrelated set.

We selected the gene-rich chromosome 19 (10756 SNPs) for analysis. Using PLINK (Purcell *et al.*, 2007) we performed linkage disequilibrium (LD) pruning on

chromosome 19 (window size = 50kb, step size = 5kb and r^2 threshold = 0.1), leaving 1705 SNPs in linkage equilibrium (LE). From the LD pruned dataset we designated the potential set of "causal" SNPs in a minor allele frequency range of 0.4 to 0.5 (584 SNPs). From this pool, high minor allele frequency SNPs were selected to ensure equally high levels of statistical power across all simulations. All analyses were performed twice: once on the linkage disequilibrium pruned (1705 SNPs) and again on the unpruned (10 756 SNPs) chromosome 19 datasets.

3.4.2 Simulation of phenotypes

We simulated phenotypes under the alternative hypothesis (H_1) and null hypothesis (H_0) with 500 replicates per condition. All added errors (ε) were drawn from a standard normal distribution N ($\mu=0$, $\sigma^2=1$). To ensure unbiased simulation, bias calculations were performed to assess possible over/under estimations of coefficients. Coverage was calculated to assess the probability that the sum of the estimated coefficients fell in the 95% confidence interval using a regression model.

Polygenic phenotype

We selected 200 SNPs from the potentially causal SNP pool to form a polygenic phenotype. In this model each SNP explains the same amount of variation ($R^2=1.5\times10^{-3}$) with a total R^2 of 0.3 (30%) (Equation 3.4). Simulations were performed using the Linkage-Disequilibrium Adjusted Kinships (LDAK) software (Speed *et al.*, 2012). Bias was observed at 0.18 and coverage was 96%. The bias calculation in the polygenic phenotype is larger compared to other phenotypes.

$$y = \beta_1 SNP_1 + \beta_2 SNP_2 + \beta_3 SNP_3 \ldots\ldots + \beta_{200} SNP_{200} + \varepsilon \quad (3.4)$$

2-SNP interacting phenotypes

Two SNPs not used for simulating the polygenic phenotype were selected at random from the potentially causal SNP pool. We simulated 2-SNP interacting

phenotypes assuming that each individual SNP has a small but present main effect ($\beta 1 \neq 0$ and $\beta 2 \neq 0$) (Equation 3.5).

$$y = \beta_1 SNP_1 + \beta_2 SNP_2 + \beta_3 SNP_1 SNP_2 + \varepsilon \quad (3.5)$$

We simulated three levels of 2-SNP interactions (weak, intermediate and strong) each explaining a different amount of variation (Table 3.1). The weak interaction phenotype strength was simulated to represent an interaction that would not be detected by a regression model after adjusting for multiple testing by means of a Bonferroni corrections (mean *p*-value = 3.1×10^{-2}; median *p*-value = 1.8×10^{-3}). The strong interaction phenotypes had a mean *p*-value = 1.3×10^{-10} and median *p*-value = 3.5×10^{-17} to assess whether c5.0 and logic regression were capable of detecting a strong signal; this phenotype was used as a proof of principle. Intermediate phenotypes (mean *p*-value = 3.6×10^{-4}; median *p*-value = 2.0×10^{-7}) were simulated to fall between the two extremes (Table 3.1). Bias calculations were all close to 0 (strong = 4.0×10^{-3}, intermediate = 6.7×10^{-4} and weak = -5.0×10^{-3}) and coverage was 96% for the strong phenotype and 94% for both the intermediate and weak phenotypes.

3-SNP interacting phenotypes

Three SNPs not previously used for simulating the polygenic phenotype were selected at random from the potentially causal SNP pool. We analysed these phenotypes independently. Three levels (weak, strong and pure) of 3-SNP interactions were simulated including all possible 2-SNP interactions (Equation 3.6).

$$y = \beta_1 SNP_1 + \beta_2 SNP_2 + \beta_3 SNP_3 + \beta_4 SNP_1 SNP_2 + \beta_5 SNP_1 SNP_3 + \beta_6 SNP_2 SNP_3 + \beta_7 SNP_1 SNP_2 SNP_3 + \varepsilon \quad (3.6)$$

We simulated a weak and strong 3-SNP interacting phenotype explaining a different amount of variation; we set the βs of the strong interaction to be twice as large as the weak ones. (Table 3.2). Also, we simulated a pure 3-SNP interaction where in Equation 3.6 β_1 to β_6 are all set to 0. Bias calculations were again all close to 0 (pure = -7.41×10^{-4}, strong = 2.26×10^{-4}, and weak = -5.87×10^{-4}) and coverage was 97% for the pure phenotype and 93% for both the intermediate and weak phenotypes.

Model	β_1,β_2	β_3	$R^2_{\text{2SNP interaction}}$ (%)	$R^2_{\text{full model}}$ (%)	Mean $p_{\text{interaction}}$	Median $p_{\text{interaction}}$
Strong	0.2	0.24	1.6	35.6	1.3×10^{-10}	3.5×10^{-17}
Intermediate	0.125	0.15	0.82	17.8	3.6×10^{-4}	2.0×10^{-7}
Weak	0.07	0.09	0.35	6.8	3.1×10^{-2}	1.8×10^{-3}

Table 3.1: Two-SNP interaction models, R^2 and *p*-values

Model	β_1,β_2,β_3	β_4,β_5,β_6	β_7	$R^2_{\text{2SNP interactions}}$ (%)	$R^2_{\text{3SNP interaction}}$ (%)	$R^2_{\text{Full model}}$ (%)	Mean $p_{\text{interaction}}$	Median $p_{\text{interaction}}$
Pure	0	0	0.4	0.04	1.86	30.1	1.0×10^{-14}	1.1×10^{-22}
Strong	0.05	0.1	0.2	0.19	0.41	39.9	4.5×10^{-4}	6.6×10^{-7}
Weak	0.025	0.05	0.1	0.15	0.10	14.3	7.7×10^{-2}	1.3×10^{-2}

Table 3.2: Three-SNP interaction models, R^2 and *p*-value

Combined polygenic and interacting phenotypes

To assess the capability of C5.0 and logic regression to detect gene-gene interactions even in the presence of an additive or polygenic component we simulated an interaction in the data used for the polygenic simulations, using SNPs not included in the polygenic component.

Null phenotype

To assess the type I error, we modelled a phenotype under H_0 where all βs are set to 0; therefore, $y = \varepsilon$. Bias was observed as 1.2×10^{-3} with a coverage of 94%.

Main effect

To rule out the possibility that the power was driven by the larger βs of the main effects within the interaction phenotypes, we simulated 500 replicates that only included the main effect of the strong phenotype; $y = \beta_1 SNP_1 + \beta_2 SNP_2 + \varepsilon$ (where β_1 and $\beta_2 = 0.2$). As this interaction has the largest coefficients we chose this setting as a proof of principle for all other phenotypes with smaller coefficients. Bias was observed as 9.0×10^{-3} and coverage of 96.4%. The strong two-SNP main effect signal was then combined with the 200-SNP polygenic signal.

3.4.3 Data pre-processing and parameter settings

Genotype data were converted to ordered vectors. Logic regression only allows for binary predictors; therefore, we dichotomised the genotype data into dominant and recessive predictors, i.e. genotype {0, 1, 2} becomes dominant {0, 1, 1} and recessive {0, 0, 1}. Missing genotypes were imputed by means of median imputation before analysis.

3.4.4 Identification of causal SNPs, and type I error and power

We defined the type I error for C5.0 as the percentage of trees constructed under H_0. Power was defined as the percentage of constructed sets of rulesets under H_1 containing all of the simulated interacting SNPs.

For logic regression, we assessed the presence of a signal under H_0 and H_1 by performing a randomisation test. Each replicate was permuted 100 times; the number of instances the original model had a lower score (residual sum of squares) than 95% of permuted models ($\alpha = 0.05$) was derived. For type I error, we counted the number of times that the replicate passed the randomisation test when no signal was present and divided by 500. For power, we considered only those replicates that passed the randomisation test, and similarly for the calculation for type I error (Figure 3.2A). Then we assessed if the logic trees contained all the simulated interacting SNPs. If the replicate (a) passed the randomisation test and (b) the simulated interacting SNPs were present, this was considered as a 'true positive' and we summed the number of these replicates and divided by 500 to obtain power (Figure 3.2B).

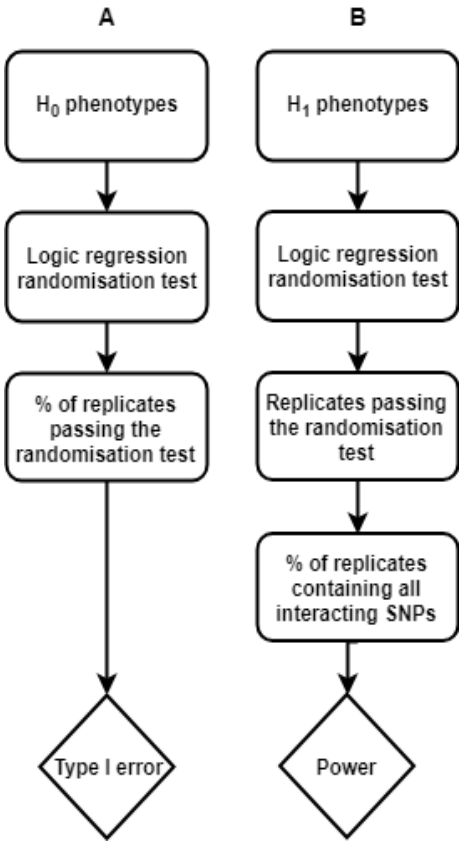


Figure 3.2: Flow chart logic regression analyses

3.5 Results

3.5.1 Type I error

We observed that C5.0 has a type I error of 0% when using the LD pruned data and 0.6% when using the LD unpruned data. Using the randomisation test we observed that logic regression has a 5.8% type I error using LD pruned data, rising to 6.4% when using LD unpruned data.

3.5.2 Power

3.5.2.1 LD pruned data

Power results for pruned and unpruned data for both methods can be found in Table 3.3. C5.0 detected rulesets in 11.4% polygenic replicates with the LD pruned dataset. Of these, 79% were based on a single SNP, 19% on two SNPs and 2% on four SNPs. All observed SNPs in the rulesets were from the 200 SNPs used to create the polygenic phenotypes, with no un-associated SNPs in any rulesets. C5.0 detected the two interacting SNPs in 100% of the strong replicates. This number decreased to 99.2% in the intermediate and 8.6% in the weak replicates (Supplementary Table S3.1). No rulesets were created that included other non-interacting SNPs; in other words, no false-positive SNPs were included in any of the rulesets generated for the interaction simulation replicates. Furthermore, in 35% of the weak replicates no ruleset was created; the remainder contained just one of the two interacting SNPs. In the combined polygenic and 2-SNP interaction phenotype analysis, C5.0 shows that it is capable of distinguishing additivity from interactions by detecting the two interacting SNPs in 100% of strong and 23% of intermediate replicates (Supplementary Table S3.1). In the combined polygenic and weak interaction analyses C5.0 did not detect a single ruleset in 98.8% of the replicates. In the remaining 1.2%, C5.0 was not able to detect both interacting SNPs (Supplementary Table S3.1). Higher order interactions, i.e. 3-SNP interaction, were also detected using C5.0. We observed a power of 100%, 100% and 90.4% of all three interacting SNPs in the pure, strong and weak 3-SNP interaction phenotypes (Supplementary Table S3.2). When combining these phenotypes with a polygenic signal, the interaction power remained 100% for the pure and strong phenotypes and dropped to

11.2% for the weak phenotype (Supplementary Table S3.2). For replicates that only contained two SNPs with main effects and no interaction, we observed that C5.0 detected 62.6% rulesets containing solely one of the two main effect SNPs, in 35.2% both SNPs and in 2.2% no rulesets. When combined with a polygenic signal this dropped to 0.4% for both SNPs and one of the main effect SNPs in 16% while it did not detect any ruleset in the remaining 84%.

Randomisation test-based analyses showed the power of logic regression using the LD pruned data ranged between 89.6% (combined polygenic and weak 3-SNP interaction) and 100% (Table 3.4) for phenotypes containing interactions and 99.6% for the polygenic-only phenotype. For each of the 500 polygenic replicates logic regression created a model containing eight SNPs. These models contained either 2 (1%), 3 (2.6%), 4 (11%), 5 (24.7%), 6 (31.1%), 7(22.9%) or 8 (6.6%) polygenic SNPs. This means that, in all replicates minus the 6.6% containing 8 polygenic SNPs, logic regression includes several SNPs that can be defined as false-positives when a polygenic signal is present because these SNPs have been LD-pruned; thus, their presence is not due to correlation with a polygenic SNP.

For all 2-SNP interaction analyses logic regression created trees containing eight SNPs with the exception of ten trees (0.15%) containing 1 (1 tree), 3 (1 tree), 4 (2 trees), 5 (2 trees), 6 (1 tree) or 7 (3 trees) SNPs. For the strong and intermediate 2-SNP phenotypes, logic regression created in 99.8% and 98.8% replicates logic trees containing both interacting SNPs (Supplementary Table S3.3). Several of the remaining SNPs in these trees were false positives, not due to LD. This dropped to 77% in the weak 2-SNP phenotype, with 1% containing no interacting SNPs. When combining the polygenic and epistatic phenotypes the trees contained the interacting SNPS in the strong (98.8%), intermediate (82.8%) and weak (9.6%) phenotypes (Supplementary Table S3.3). Furthermore, 45.3% of the created combined polygenic-weak trees contained no interacting SNPs. The majority of trees in the higher order 3-SNP interaction analyses contained the interacting SNPs (pure 99.6%, strong 99.6% and weak 89%; Supplementary Table S3.4A). This number decreased when adding the polygenic component (pure 98.4%, strong 95.2% and weak 53.6%; Supplementary Table S3.4B). No trees were observed containing solely non interacting SNPs.

3.5.2.2 Unpruned data

C5.0 detection of rulesets in the polygenic model increased to 53.4% when not LD pruned (Supplementary Table S3.5). Seventy-six point eight percent of observed SNPs in the rulesets were used to create the 200 SNP polygenic phenotypes or were in LD with a polygenic SNP (r^2 >0.25). Compared to the pruned set analyses, the percentage accurately detecting 2-SNP interactions by C5.0 remained 100% for the strong phenotype, but decreased to 98.2% in the intermediate phenotype. The percentage accurately detected 2-SNP interactions increased to 19.8% in the weak phenotype (Supplementary Table S3.6). However, it has to be noted that C5.0 detected in 3.2% (16 rulesets) non-interacting random SNPs of which 12 contained SNPs in LD with the true signal (r^2 > 0.25; this threshold was set to be consistent with the value for LD pruning). In the combined polygenic and interaction phenotype analysis, the power remained unchanged for the strong phenotype. The power was again higher in the intermediate (41.6%) and weak (0.6%) phenotype, but 10.6% and 6.6%, of replicates respectively, contained at least one false positive SNP (Supplementary Table S3.6), which could be linked to LD structure. We observed no change in C5.0 interaction power in the pure and strong three 3-SNP interaction phenotypes (100%) and an increase to 91% in the weak phenotype (Supplementary Table S3.7). Only the weak interaction phenotype showed a higher power compared to the pruned analysis of 24% (Supplementary Table S3.7).

Randomisation test-based analyses using the non-LD pruned data showed six interaction phenotypes having a lower power when using non-LD pruned data compared to LD pruned data. Power dropped to 94.4% for the polygenic phenotype. The largest differences were observed with the combined polygenic and 3-SNP interaction phenotype (35.6%) and weak 2-SNP phenotype (32.8%) (Table 3.3). When analysing the polygenic phenotype we observed that logic regression created 15.4% trees containing no polygenic SNPs. This dropped to 1.4% when taking LD structure into account (r^2 >0.25). The rest of the trees contained either 1 (33.4%), 2 (29.6%), 3 (14.6%), 4 (5.6%), 5 (1%) or 6 (0.4%) polygenic SNPS and in 89.3% in combination with numerous SNPs in LD with the polygenic SNPs. The power of logic regression for the two interacting SNP phenotypes was 52.8% in the strong, 17.6% in the intermediate and 0.3% in the weak phenotype (Supplementary Table

S3.8). This dropped further in the combined analysis to 23.4% in the strong, 2.2% in the intermediate and 0% in the weak phenotype (Supplementary Table S3.8).

We observed that in the higher order phenotypes, the trees contain three forms of the interacting SNPs is 50.2% (pure), 35.4% (strong) and 14.7% (weak) (Supplementary Table S9a). In line with previous observed results when adding the polygenic signal the numbers again lowered to 35.2% (pure), 21% (strong) and 2.2% (weak) (Supplementary Table S9b). For all phenotypes a percentage of trees were created containing non interacting SNPs; however, the majority of these trees contained SNPs in LD (r^2 >0.25) with the interacting SNPs (for a detailed outline see Supplementary Table S3.10).

	Power	
Model	**Pruned**	**Unpruned**
Weak 2-SNP interaction	97.4	64.6
Intermediate 2-SNP interaction	100	100
Strong 2-SNP interaction	100	100
30% Polygenic + Weak 2-SNP interaction	89.6	54
30% Polygenic + Intermediate 2-SNP interaction	100	89.6
30% Polygenic + Strong 2-SNP interaction	100	100
Weak 3-SNP interaction	100	99.2
Strong 3-SNP interaction	100	100
Pure 3-SNP interaction	100	100
30% Polygenic + Weak 3-SNP interaction	100	92.6
Polygenic + Strong 3-SNP interaction	100	100
Polygenic + Pure 3-SNP interaction	100	100
Polygenic	99.6	94.4

Table 3.3: Power of logic regression, based on randomisation tests

3.6 Application to educational attainment in GS:SFHS

Having assessed our methods by simulation, we wished to test the approach on a large set of complex trait data. We extracted 7,012 unrelated GS:SFHS individuals of which 6,765 individuals had a measure of years of education, measured by ordered categories (e.g. 0: 0 years, 1: 1-4 years, 2: 5-9 years). We performed a linear regression analysis between years of education controlling for sex and age, and extracted the residuals to act as an adjusted years of education measurement (Zhao *et al.*, 2012).

Finally, we applied C5.0 and logic regression on the residual years of education outcome using 131,821 whole genome SNPs in LE (LD pruning settings; window size = 50kb, step size = 5kb and r^2 threshold = 0.1). C5.0 detected 32 rulesets associated with educational attainment containing in total 30 SNPs (Supplementary Table S3.11). The logic regression model did not pass the randomisation test (α=0.05) so will not be discussed further.

3.7 Conclusions and Discussion

When using LD-pruned genetic data we observed that C5.0 is capable of distinguishing additivity from interactions. C5.0 created rulesets based on a polygenic phenotype in 11.4% of the replicates; however, the majority of these (78.9%) were based on one single polygenic SNP. C5.0 correctly detected both interacting SNPs in 100 and 99.2% in the strong and intermediate phenotypes. Even though the interaction strength was low, C5.0 was capable of detecting the signal in 8.6% of the weak 2-SNP interaction replicates, of which none would be significant using a standard regression model after adjusting for multiple testing. For the 3-SNP (higher order) interaction phenotype, C5.0 was able to detect all three SNPs in 100% of the pure and strong and in 90.4% of the weak phenotype. When combining the polygenic and interaction phenotypes C5.0 was able to distinguish the interaction signal from the polygenic signal in 100 and 23% of the strong and intermediate 2-SNP phenotypes. For the weak phenotype C5.0, was not able to detect any ruleset in 98.8% of the replicates showing it to be protective against spurious results when the interaction term is of low magnitude. Similar results were observed in the 3-SNP combined analyses. As no rulesets were observed under H0, we conclude that C5.0 had a low type I error. We could not see any evidence that our previously observed results were driven by main effects when analyzing strong main effect data only. This indicates that C5.0 is detecting rulesets based on conditional dependencies and not on large main effects. We observed that LD structure has an impact on the performance of C5.0. In all but four phenotypes that include an interaction component the amount of accurately detected interactions decreased using unpruned data (Table 3.4).

We observed that logic regression is capable of accurately detecting all

interacting SNPs in all but one phenotype either combined with additivity and using LD pruned or unpruned data. Logic regression was not capable of detecting both interacting SNPs in the 2-SNP interaction including a polygenic signal in the LD unpruned phenotype. However, we observed a slightly inflated type I error (5.8 and 6.8%), which is in line with the developers' statement that logic regression is likely to overfit (Kooperberg *et al.*, 2001). It should be noted that logic regression has a high overall power when performing a randomisation analysis, however when looking into the SNPs used to create the initial model, logic regression-built trees using random SNPs therefore the overall randomisation test-based power is high but the frequency of inclusion of spurious SNPs in a model is also high. Furthermore, as mentioned logic regression applies a greedy hill climbing algorithm. Greedy hill climbing algorithms stop when the last predictor included does not improve the prediction rate. As logic regression applies a random starting point, it risks creating a set of Boolean combinations of binary predictors that may reflect a local optimum rather than the global optimum. One solution to circumvent this issue is to apply a global optimum search technique e.g. simulated annealing.

			Power			
Interaction	**Data**	**Strength**	**Without polygenic phenotype**		**Combined with polygenic phenotype**	
			C5.0	**Logic regression**	**C5.0**	**Logic regression**
2-SNP	Pruned	Weak	8.6%	77%	**0%**	9.6%
		Intermediate	99.2%	98.8%	23%	82.8%
		Strong	100%	99.8%	100%	98.8%
	Unpruned	Weak	19.8%	0.3%	0.6%	**0%**
		Intermediate	98.2%	17.6%	41.6%	2.2%
		Strong	100%	52.8%	100%	23.4%
3-SNP	Pruned	Weak	90.4%	89%	3.6%	53.6%
		Strong	100%	99.6%	100%	95.2%
		Pure	100%	99.6%	100%	98.4%
	Unpruned	Weak	91%	14.7%	24%	2.2%
		Strong	100%	35.4%	100%	21%
		Pure	100%	50.2%	100%	35.2%

Table 3.4: Power of C5.0 and logic regression in pruned and unpruned data, with and without polygenic signal

We observed 32 rulesets containing 30 putative epistatic SNPs associated with educational attainment (EA) in Generation Scotland. From the thirty SNPs, 18 could be mapped to genes, two were in genes previously associated with mental health or cognitive performance (rs196433, chr1, *RCAN3* and rs17100828, chr14, *NPAS3*). *RCAN3* is associated with reading and spelling (Luciano *et al.*, 2013) while *NPAS3* acts as a master regulator of neuropsychiatric risk genes (Michaelson *et al.*, 2017). Of the remaining 16 genes none showed a clear association with any phenotype. We sought to investigate whether these SNPs have been previously reported in the large EA GWAS study which observed 74 statistically significant SNPs (Okbay *et al.*, 2016). None of the SNPs observed in this study overlapped or could be considered a proxy SNP ($r^2 > 0.8$) with the previously reported GWAS results. One explanation for the lack of overlap might be because GWAS searches for single SNPs associated with a phenotype while C5.0 searches for conditional dependencies associated with the phenotype. Therefore one could say that both methods search for different pieces of the same puzzle. The results strengthens the assumption that interacting SNPs play an important role in educational attainment (Supplementary Table S3.12).

The main strength of this study is that we assessed the capability of both C5.0 and logic regression in detecting simulated genetic interactions under a wide range of settings including a strong polygenic signal. We suggest that C5.0 rulesets might be used as predictors within a regression model alongside single SNPs and additive or polygenic components (Nicodemus *et al.*, 2014). The same can be done with logic trees. Limitations lie in the modest sample size (n=5,000) and the use of only causal SNPs with a large MAF (0.4-0.5). We did not simulate phenotypes containing multiple SNP interactions (polygenic-epistatic phenotype) which is biologically plausible.

In conclusion, we have shown that C5.0 and logic regression are capable of detecting simulated genetic interactions in a wide range of association levels and even in the presence of a strong polygenic component. We showed that when applying both methods LD pruning helps by improving the power and reducing the type I error. Finally, using C5.0 we were able to detect 32 rulesets containing 30 SNPs not

previously reported with EA in Generation Scotland; RCAN3 has been previously observed in association with learning and reading while NPAS3 is involved in neurodevelopment. These methods are capable of detecting SNPs not directly associated to the trait but rather in sets of SNPs that together affect the trait. These methods are well-adapted to testing hypotheses regarding the 'omnigenic' model.

Brief conclusion

This published work is a proof of principle to show that both tree based techniques are able to detect epistasis even in the presence of a phenotype containing strong-single locus and polygenic components. Both tree based methods used in this published work can now be used alongside methodologies described in chapter 2 (single loci and polygenic scores) thereby extending the work laid out in Nicodemus *et al*., (2008).

3.8 References

Bateson, W. and Mendel, G. (1909) 'Mendel's Principles of heredity, by W. Bateson'. Cambridge: University Press, pp. 1–450. doi: 10.5962/bhl.title.44575.

Boyle, E. A., Li, Y. I. and Pritchard, J. K. (2017) 'An Expanded View of Complex Traits: From Polygenic to Omnigenic', *Cell*, pp. 1177–1186. doi: 10.1016/j.cell.2017.05.038.

Brockmann, G. A. *et al.* (2000) 'Single QTL effects, epistasis, and pleiotropy account for two-thirds of the phenotypic F2 variance of growth and obesity in DU6i x DBA/2 mice', *Genome Research*. Cold Spring Harbor Laboratory Press, 10(12), pp. 1941–1957. doi: 10.1101/gr.GR1499R.

Calvin, C. M. *et al.* (2012) 'Multivariate genetic analyses of cognition and academic achievement from two population samples of 174,000 and 166,000 school children', *Behavior Genetics*, 42(5), pp. 699–710. doi: 10.1007/s10519-012-9549-7.

Chen, C. C. M. *et al.* (2011) 'Methods for identifying SNP interactions: A review on variations of logic regression, Random Forest and bayesian logistic regression', *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pp. 1580–1591. doi: 10.1109/TCBB.2011.46.

Cheng, Y. *et al.* (2011) 'Mapping genetic loci that interact with myostatin to affect growth traits', *Heredity*. Nature Publishing Group, 107(6), pp. 565–573. doi: 10.1038/hdy.2011.45.

Fisher, R. A. (1919) 'XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance.', *Transactions of the Royal Society of Edinburgh*, 52(2), pp. 399–433. doi: 10.1017/S0080456800012163.

García-Magariños, M. *et al.* (2009) 'Evaluating the ability of tree-based methods and logistic regression for the detection of SNP-SNP interaction', *Annals of Human Genetics*, 73(3), pp. 360–369. doi: 10.1111/j.1469-1809.2009.00511.x.

Grice, S. J., Liu, J. L. and Webber, C. (2015) 'Synergistic Interactions between Drosophila Orthologues of Genes Spanned by De Novo Human CNVs Support

Multiple-Hit Models of Autism', *PLoS Genetics*. Edited by S. Shifman, 11(3), p. e1004998. doi: 10.1371/journal.pgen.1004998.

He, X. *et al.* (2016) 'Epistatic partners of neurogenic genes modulate Drosophila olfactory behavior', *Genes, Brain and Behavior*, 15(2), pp. 280–290. doi: 10.1111/gbb.12279.

Hill, W. G., Goddard, M. E. and Visscher, P. M. (2008) 'Data and theory point to mainly additive genetic variance for complex traits', *PLoS Genetics*, 4(2). doi: 10.1371/journal.pgen.1000008.

Huang, W. *et al.* (2012) 'Epistasis dominates the genetic architecture of Drosophila quantitative traits', *Proceedings of the National Academy of Sciences*, 109(39), pp. 15553–15559. doi: 10.1073/pnas.1213423109.

Huang, W. and Mackay, T. F. C. (2016) 'The Genetic Architecture of Quantitative Traits Cannot Be Inferred from Variance Component Analysis', *PLoS Genetics*. Edited by X. Zhu, 12(11), p. e1006421. doi: 10.1371/journal.pgen.1006421.

Jang, K. L., Livesley, W. J. and Vemon, P. A. (1996) 'Heritability of the Big Five Personality Dimensions and Their Facets: A Twin Study', *Journal of Personality*, 64(3), pp. 577–592. doi: 10.1111/j.1467-6494.1996.tb00522.x.

Kooperberg, C. *et al.* (2001) 'Sequence analysis using logic regression.', *Genetic epidemiology*, 21 Suppl 1(Suppl 1), pp. S626-31.

Krapohl, E. *et al.* (2014) 'The high heritability of educational achievement reflects many genetically influenced traits, not just intelligence', *Proceedings of the National Academy of Sciences*, 111(42), pp. 15273–15278. doi: 10.1073/pnas.1408777111.

Kuhn, M. and Johnson, K. (2013) 'Introduction', in *Applied Predictive Modeling*. New York, NY: Springer New York, pp. 1–16. doi: 10.1007/978-1-4614-6849-3_1.

Loehlin, J. C., Neiderhiser, J. M. and Reiss, D. (2003) 'The behavior genetics of personality and the NEAD study', *Journal of Research in Personality*. Academic Press, 37(5), pp. 373–387. doi: 10.1016/S0092-6566(03)00012-6.

Lu, Q. *et al.* (2012) ‘A Likelihood Ratio-Based Mann-Whitney Approach Finds Novel Replicable Joint Gene Action for Type 2 Diabetes’, *Genetic Epidemiology*, 36(6), pp. 583–593. doi: 10.1002/gepi.21651.

Luciano, M. *et al.* (2013) ‘A genome-wide association study for reading and language abilities in two population cohorts’, *Genes, Brain and Behavior*. Blackwell Publishing Ltd, 12(6), pp. 645–652. doi: 10.1111/gbb.12053.

Mackay, T. F. C. (2014) ‘Epistasis and quantitative traits: using model organisms to study gene-gene interactions.’, *Nature reviews. Genetics*, 15(1), pp. 22–33. doi: 10.1038/nrg3627.

Mackay, T. F. and Moore, J. H. (2014) ‘Why epistasis is important for tackling complex human disease genetics’, *Genome Medicine*. BioMed Central, 6(6), p. 125. doi: 10.1186/gm561.

Michaelson, J. J. *et al.* (2017) ‘Neuronal PAS Domain Proteins 1 and 3 Are Master Regulators of Neuropsychiatric Risk Genes’, *Biological Psychiatry*, 82(3), pp. 213–223. doi: 10.1016/j.biopsych.2017.03.021.

de Moor, M. H. M. *et al.* (2012) ‘Meta-analysis of genome-wide association studies for personality’, *Molecular Psychiatry*, 17(3), pp. 337–349. doi: 10.1038/mp.2010.128.

Nicodemus, K. K. *et al.* (2014) ‘Variability in Working Memory Performance Explained by Epistasis vs Polygenic Scores in the ZNF804A Pathway.’, *JAMA Psychiatry*, 71, pp. 778–785. doi: 10.1001/jamapsychiatry.2014.528.

Okbay, A. *et al.* (2016) ‘Genome-wide association study identifies 74 loci associated with educational attainment’, *Nature*. Nature Publishing Group, 533(7604), pp. 539–542. doi: 10.1038/nature17671.

Power, R. A. and Pluess, M. (2015) ‘Heritability estimates of the Big Five personality traits based on common genetic variants’, *Translational Psychiatry*, 5(7). doi: 10.1038/tp.2015.96.

Purcell, S. *et al.* (2007) ‘PLINK: A Tool Set for Whole-Genome Association and

Population-Based Linkage Analyses', *The American Journal of Human Genetics*, 81(3), pp. 559–575. doi: 10.1086/519795.

Quinlan, J. R. (1992) *C4.5: Programs for Machine Learning, Morgan Kaufmann San Mateo California*. Morgan Kaufmann Publishers. doi: 10.1016/S0019-9958(62)90649-6.

Rietveld, C. A. *et al.* (2013) 'GWAS of 126,559 individuals identifies genetic variants associated with educational attainment', *Science*, 340(6139), pp. 1467–1471. doi: 10.1126/science.1235488.

Ruczinski, I., Kooperberg, C. and LeBlanc, M. (2003) 'Logic Regression', *Journal of Computational and Graphical Statistics*, 12(3), pp. 475–511. doi: 10.1198/1061860032238.

Ruczinski, I., Kooperberg, C. and LeBlanc, M. L. (2004) 'Exploring interactions in high-dimensional genomic data: An overview of Logic Regression, with applications', *Journal of Multivariate Analysis*, pp. 178–195. doi: 10.1016/j.jmva.2004.02.010.

Sackton, T. B. and Hartl, D. L. (2016) 'Genotypic Context and Epistasis in Individuals and Populations', *Cell*, pp. 279–287. doi: 10.1016/j.cell.2016.06.047.

Schwender, H. and Ickstadt, K. (2008) 'Identification of SNP interactions using logic regression', *Biostatistics*, 9(1), pp. 187–198. doi: 10.1093/biostatistics/kxm024.

Smith, B. H. *et al.* (2006) 'Generation Scotland: the Scottish Family Health Study; a new resource for researching genes and heritability.', *BMC medical genetics*, 7(1), p. 74. doi: 10.1186/1471-2350-7-74.

Smith, B. H. *et al.* (2013) 'Cohort profile: Generation scotland: Scottish family health study (GS: SFHS). The study, its participants and their potential for genetic research on health and illness', *International Journal of Epidemiology*, 42(3), pp. 689–700. doi: 10.1093/ije/dys084.

Speed, D. *et al.* (2012) 'Improved heritability estimation from genome-wide SNPs', *American Journal of Human Genetics*. Elsevier, 91(6), pp. 1011–1021. doi:

10.1016/j.ajhg.2012.10.010.

Stephan, J., Stegle, O. and Beyer, A. (2015) ‘A random forest approach to capture genetic effects in the presence of population structure’, *Nature Communications*, 6. doi: 10.1038/ncomms8432.

Vukasović, T. and Bratko, D. (2015) ‘Heritability of personality: A meta-analysis of behavior genetic studies’, *Psychological Bulletin*, 141(4), pp. 769–785. doi: 10.1037/bul0000017.

Webber, C. (2017) ‘Epistasis in Neuropsychiatric Disorders’, *Trends in Genetics*, pp. 256–265. doi: 10.1016/j.tig.2017.01.009.

Wu, X. *et al.* (2008) ‘Top 10 algorithms in data mining’, *Knowledge and Information Systems*, 14(1), pp. 1–37. doi: 10.1007/s10115-007-0114-2.

Yang, J. *et al.* (2011) ‘GCTA: a tool for genome-wide complex trait analysis.’, *American journal of human genetics*, 88(1), pp. 76–82. doi: 10.1016/j.ajhg.2010.11.011.

Zhao, Y. *et al.* (2012) ‘Correction for population stratification in random forest analysis’, *International Journal of Epidemiology*. Oxford University Press, 41(6), pp. 1798–1806. doi: 10.1093/ije/dys183.

3.9 Supplementary material

Phenotypic model	%	No rulesets	Two interacting SNPs	One interacting SNP	Non-interacting SNPs
Strong 2-SNP interaction	100		✓		
Intermediate 2-SNP interaction	99.2		✓		
	0.8			✓	
Weak 2-SNP interaction	8.6		✓		
	56.4			✓	
	35	✓			
30% polygenic + strong 2-SNP interaction	99.6		✓		
	0.4		✓		✓
30% polygenic + Intermediate 2-SNP interaction	66			✓	
	0.2			✓	✓
	23		✓		
	10.8	✓			
30% polygenic + Weak 2-SNP interaction	0.8			✓	
	0.2				✓
	0.2			✓	✓
	98.8	✓			

Supplementary Table S3.1: C5.0 results in 2-SNP interaction phenotypes LD pruned data.

Phenotypic model	%	No rulesets	Three interacting SNPs	Two interacting SNPs	One interacting SNP	Non-interacting SNPs
Pure 3-SNP interaction	99.6		✓			
	0.4		✓			✓
Strong 3-SNP interaction	100		✓			
Weak3-SNP interaction	0.2				✓	
	9.4			✓		
	90.2		✓			
	0.2		✓			✓
30% polygenic + pure 3-SNP interaction	98.4		✓			
	1.6		✓			✓
Polygenic + strong 3-SNP interaction	98.2		✓			
	1.8		✓			✓
Polygenic + Weak 3-SNP interaction	10.8	✓				
	39.4				✓	
	0.2				✓	✓
	0.2			✓		✓
	38.2			✓		
	10.6		✓			
	0.6		✓			✓

Supplementary Table S3.2: C5.0 results in 3-SNP interaction phenotypes LD pruned data.

Phenotypic model	Without polygenic phenotype			Combined with polygenic phenotype		
	% SNP_1	% SNP_2	% SNP_1 SNP_2	% SNP_1	% SNP_2	% SNP_1 SNP_2
Weak 2-SNP interaction	14	8	77	35.3	10.3	9.6
Intermediate 2-SNP interaction	1	0.2	98.8	15.4	1.8	82.8
Strong 2-SNP interaction	0.2	0	99.8	1.2	0	98.8

Supplementary Table S3.3: Logic regression results in 2-SNP interaction phenotypes LD pruned data.

Phenotypic model	Without polygenic phenotype						
	% SNP_1	% SNP_2	% SNP_3	% SNP_1 SNP_2	% SNP_1 SNP_3	% SNP_2 SNP_3	% SNP_1 SNP_2 SNP_3
Weak 3-SNP interaction	0	0	0	2.2	2.6	6.2	89
Strong 3-SNP interaction	0	0	0	0	0	0.4	99.6
Pure 3-SNP interaction	0	0	0	0.2	0	0.2	99.6

Supplementary Table S3.4A: Logic regression results in 3-SNP interaction phenotypes LD pruned data.

Phenotypic model	Combined with polygenic phenotype						
	% SNP_1	% SNP_2	% SNP_3	% SNP_1 SNP_2	% SNP_1 SNP_3	% SNP_2 SNP_3	% SNP_1 SNP_2 SNP_3
Weak 3-SNP interaction	5.6	2.8	1.6	19	7.2	10.2	53.6
Strong 3-SNP interaction	0	0	0	1.2	0.8	2.8	95.2
Pure 3-SNP interaction	0	0	0	0.6	0.2	0	98.4

Supplementary Table S3.4B: Logic regression results in 3-SNP interaction phenotypes LD pruned data.

Ruleset size (*n* SNPS)	Rounded % rulesets (*n* total = 233)
1	39
2	25
3	12
4	9
5	4
6	4
7	2
8	3
10	0.4
11	0.4
12	0.4

Supplementary Table S3.5: Detailed outline of the distribution of decision trees created by C5.0 using the polygenic phenotype.

Phenotypic model	%	No rulesets	Two interacting SNPs	One interacting SNP	Non-interacting SNPs
Strong 2-SNP interaction	96.8		✓		
	3.2		✓		✓
Intermediate 2-SNP interaction	1.8			✓	✓
	89.4		✓		
	8.8		✓		✓
Weak 2-SNP interaction	23.6	✓			
	3.2				✓
	51.8			✓	
	1.6			✓	✓
	18.4		✓		
	1.4		✓		✓
30% Polygenic + Strong 2-SNP interaction	85.4		✓		
	14.6		✓		✓
30% Polygenic + Intermediate 2-SNP interaction	48.6			✓	
	1.8				✓
	35		✓		
	3.8			✓	✓
	6.6		✓		✓
	4.2	✓			
30% Polygenic + Weak 2-SNP interaction	86.6	✓			
	4.2			✓	
	6				✓
	0.2		✓		
	2.6			✓	✓
	0.4		✓		✓

Supplementary Table S3.6: C5.0 results in 2-SNP interaction phenotypes non LD pruned data

Phenotypic model	%	No rulesets	Three interacting SNPs	Two interacting SNPs	One interacting SNP	Non-interacting SNPs
Pure 3-SNP interaction	90		✓			
	10		✓			✓
Strong 3-SNP interaction	88.6		✓			
	11.4		✓			✓
Weak 3-SNP interaction	0.4				✓	
	0.2				✓	✓
	4.4			✓		
	4			✓		✓
	82.8		✓			
	8.2		✓			✓
30% Polygenic + pure 3-SNP interaction	80.8		✓			
	19.2		✓			✓
30% Polygenic + strong 3-SNP interaction	82.2		✓			
	17.8		✓			✓
30% Polygenic + Weak 3-SNP interaction	22.6				✓	
	0.4					✓
	39.4			✓		
	3				✓	✓
	18.2		✓			
	6.2			✓		✓
	5.8		✓			✓
	2.4	✓				

Supplementary Table S3.7: C5.0 results in 3-SNP interaction phenotypes non LD pruned data.

	Without polygenic phenotype			Combined with polygenic phenotype		
Phenotypic model	**% SNP_1**	**% SNP_2**	**% SNP_1 SNP_2**	**% SNP_1**	**% SNP_2**	**% SNP_1 SNP_2**
Weak 2-SNP interaction	14.6	9.9	0.3	7.2	0.6	0
Intermediate 2-SNP interaction	25.6	31.8	17.6	21.4	7.4	2.2
Strong 2-SNP interaction	21.2	22.2	52.8	33.8	23.6	23.4

Supplementary Table S3.8: Logic regression results in 2-SNP interaction phenotypes non LD pruned data.

	Without polygenic phenotype						
Phenotypic model	**% SNP_1**	**% SNP_2**	**% SNP_3**	**% SNP_1 SNP_2**	**% SNP_1 SNP_3**	**% SNP_2 SNP_3**	**% SNP_1 SNP_2 SNP_3**
Weak 3-SNP interaction	19.6	4.8	11	17.1	17.7	10.3	14.7
Strong 3-SNP interaction	4.8	1.4	9.6	5.4	34.2	9.2	35.4
Pure 3-SNP interaction	3	0	4.4	2.4	31.4	8.6	50.2

Supplementary Table S3.9A: Logic regression results in 3-SNP interaction phenotypes non LD pruned data.

	Combined with polygenic phenotype						
Phenotypic model	**% SNP_1**	**% SNP_2**	**% SNP_3**	**% SNP_1 SNP_2**	**% SNP_1 SNP_3**	**% SNP_2 SNP_3**	**% SNP_1 SNP_2 SNP_3**
Weak 3-SNP interaction	20.7	11.2	13.6	8.6	6.5	3.5	2.2
Strong 3-SNP interaction	20.4	1.6	5.6	14	29	8.4	21
Pure 3-SNP interaction	11.8	0.8	3.8	12.4	26	9	35.2

Supplementary Table S3.9B: Logic regression results in 3-SNP interaction phenotypes non LD pruned data.

Phenotype	% non-interacting SNPs	% in LD with all interacting SNPs*
Weak 2-SNP	75.2	6.8
Intermediate 2-SNP	25	20.4
Strong 2-SNP	3.8	84.2
30% Polygenic + Weak 2-SNP	91.9	0.7
30% Polygenic + Intermediate 2-SNP	69.6	10.8
30% Polygenic + Strong 2-SNP	19.2	18.3
Weak 3-SNP	4.8	0
Strong 3-SNP	0	0
Extreme 3-SNP	0	0
30% Polygenic + Weak 3-SNP	33	0
30% Polygenic Strong 3-SNP	0	0
30% Polygenic Extreme 3-SNP	0.8	0

Supplementary Table S3.10: Percentage of logic trees containing non-interacting SNPs and, of those the percentage in LD with all interacting SNPs.
**this count is the number of trees that contained SNPs that were in LD with all of the SNPs that interacted (e.g., either 2 SNPs or 3 SNPs, depending on the simulation model)*

Ruleset number	Ruleset build	Outcome	N individuals	Ruleset number	Ruleset build	Outcome	N individuals
1	rs196433 > 0 rs6747637 > 0 rs10125618 = 2 rs7965873 > 0 rs7226712 > 0 rs7256201 = 0	-0.9582104	33	17	rs4416197 = 0 rs6747637 > 0 rs4279287 = 0 rs10216277 > 0 rs7965873 = 0 rs17100828 < 2 rs12923539 < 2	0.6721485	144
2	rs7567614 = 0 rs17057882 < 2 rs10993564 = 2 rs11878345 > 0	-0.9667753	101	18	rs196433 > 0 rs7567614 = 0 rs6747637 > 0 rs10125618 < 2 rs3802609 < 2 rs7965873 > 0 rs1442849 > 0 rs7226712 > 0	0.4464759	221
3	rs6747637 > 0 rs4279287 = 0 rs10216277 = 0 rs4739619 < 2 rs1403257 < 2 rs7965873 = 0 rs12923539 < 2 rs7226712 > 0	-0.7254973	883	19	rs7567614 > 0 rs477995 > 0	0.859727	212
4	rs196433 > 0	-0.6437932	237	20	rs7567614 = 0	0.8822943	253

	rs7567614 = 0 rs6747637 > 0 rs10125618 < 2 rs3802609 < 2 rs7965873 > 0 rs282593 < 2 rs1442849 = 0 rs7226712 > 0				rs6747637 > 0 rs2404867 > 0 rs17057882 < 2 rs10993564 < 2 rs7226712 = 0 rs11878345 > 0		
5	rs7567614 = 0 rs4416197 = 0 rs6747637 > 0 rs1403257 = 2 rs13334339 > 0 rs12923539 < 2 rs7226712 > 0 rs2206173 < 2	-0.801764	126	21	rs3770613 = 2 rs4416197 = 0 rs7965873 = 0 rs12923539 = 2 rs7226712 > 0	1.3422044	84
6	rs7567614 = 0 rs4416197 > 0 rs6747637 > 0 rs7965873 = 0 rs7226712 > 0	-0.6803628	854	22	rs7567614 = 0 rs4416197 = 0 rs6747637 > 0 rs4739619 = 2 rs7965873 = 0 rs17100828 < 2 rs12923539 < 2 rs7226712 > 0	1.04883	37
7	rs7567614 = 0	-0.6437932	1365	23	rs4416197 = 0	1.4013413	33

	rs6747637 = 0 rs10934116 = 0 rs4947631 = 0				rs6747637 > 0 rs10216277 = 0 rs1403257 = 2 rs7965873 = 0 rs13334339 = 0 rs7226712 > 0 rs2206173 < 2		
8	rs7567614 = 0 rs10934116 = 0 rs4947631 > 0 rs11054372 = 0	-0.5760914	107	24	rs7567614 = 0 rs4416197 = 0 rs6747637 > 0 rs7965873 = 0 rs12923539 = 2 rs765742 = 1	1.4916104	73
9	rs196433 = 0 rs7567614 = 0 rs6747637 > 0 rs7965873 > 0 rs17100828 < 2 rs7226712 > 0	-0.5900937	219	25	rs196433 > 0 rs6747637 > 0 rs10125618 < 2 rs7965873 > 0 rs282593 = 2	2.4690432	35
10	rs3770613 < 2 rs4416197 = 0 rs6747637 > 0 rs7965873 = 0 rs12923539 = 2 rs7226712 > 0	-0.4632551	167	26	rs7567614 = 0 rs10125618 < 2 rs3802609 = 2 rs7965873 > 0 rs7990443 = 2 rs7226712 > 0	1.7468907	53

	rs765742 = 0						
11	rs7567614 > 0 rs477995 = 0	-0.463255	161	27	rs6747637 > 0 rs3802609 = 2 rs7965873 > 0 rs282593 < 2 rs7226712 > 0 rs2206173 = 0	1.6495813	61
12	rs7567614 = 0 rs6747637 > 0 rs10125618 < 2 rs3802609 = 2 rs7965873 > 0 rs7990443 < 2 rs282593 < 2 rs2206173 > 0	-0.2531093	167	28	rs4416197 = 0 rs6747637 > 0 rs4279287 = 0 rs10216277 = 0 rs4739619 < 2 rs1403257 = 2 rs7965873 = 0 rs12923539 < 2 rs7226712 > 0 rs2206173 = 2	1.3787741	30
13	rs7567614 = 0 rs6747637 > 0 rs17057882 < 2 rs17100828 < 2 rs7226712 = 0 rs11878345 = 0	-0.372986	727	29	rs6747637 > 0 rs17057882 = 2 rs7226712 = 0	1.3562067	53
14	rs2404867 = 0 rs17057882 < 2	-0.0725712	779	30	rs17100828 = 2	1.5453099	24

	rs10993564 < 2 rs17100828 < 2 rs7226712 = 0 rs11878345 > 0						
15	rs7567614 = 0 rs6747637 = 0 rs10934116 > 0	-0.2756765	604	31	rs196433 > 0 rs7567614 = 0 rs6747637 > 0 rs10125618 = 2 rs7965873 > 0 rs7226712 > 0 rs7256201 > 0	2.0628324	45
16	rs7567614 = 0 rs4279287 > 0 rs4739619 < 2 rs7965873 = 0 rs7226712 > 0	0.4550408	297	32	rs7567614 = 0 rs6747637 = 0 rs10934116 = 0 rs4947631 > 0 rs11054372 > 0	2.5001753	22

Supplementary Table S3.11: C5.0 rulesets associated with educational attainment.

Marker Name	CHR	Pos	P
rs196433	1	24862643	0.9484
rs3770613	2	170146913	0.03686
rs4416197	2	201596926	0.3345
rs6747637	2	212406789	0.4191
rs7567614	2	54386795	0.9382
rs10934116	3	111574943	0.1779
rs2404867	4	136519891	0.5269
rs4279287	4	189151984	0.9237
rs17057882	5	159926960	0.04366
rs10216277	7	4401148	0.05367
rs4947631	7	50603379	0.6259
rs4739619	8	82149338	0.2022
rs10125618	9	6555311	0.002604
rs10993564	9	93272277	0.3226
rs3802609	10	27025659	0.6269
rs477995	11	104737590	0.0455
rs1403257	11	80789029	0.5751
rs7965873	12	44354797	0.1481
rs11054372	12	11757164	0.6413
rs7990443	13	79494019	0.584
rs282593	13	113378882	0.5949
rs17100828	14	33839502	0.3818
rs13334339	16	8615618	0.2977
rs12923539	16	59582084	0.4694
rs1442849	17	8024121	0.05291
rs7226712	18	2874552	0.1987
rs765742	19	57151907	0.07446
rs7256201	19	51705039	0.1543
rs11878345	19	35868773	0.7682
rs2206173	22	35223541	0.4449

Supplementary Table S3.12: p-values of SNPs observed in C5.0 rulesets in *Okbay et al, 2016*.

4 Combining single-loci, polygenic risk scores and SNP-SNP interactions to explain a significant proportion of variation in neuroticism

4.1 Introduction

Neuroticism is a moderately stable personality trait characterised as a tendency to respond with a negative emotional response to threat, frustration, or loss (Matthews, Deary and Whiteman, 2009). Higher levels of neuroticism are associated with poorer mental and physical health, making it a well-defined risk factor for negative health outcomes (Lahey, 2009). The direct and indirect financial burden of neuroticism on a society is significant. The cost of neuroticism per 1 million individuals falling in the top 25% of neuroticism levels was estimated to be around $1.393 billion per year in the Netherlands (Cuijpers *et al.*, 2010). Genetic contributions to neuroticism have been established. In the most recent genomic investigation of neuroticism, using a large (*n* = 329,821) Genome-Wide Association Study (GWAS), 116 loci were detected that were genome-wide significantly associated with neuroticism scores (Luciano *et al.*, 2018). Polygenic scores derived using summary statistics of the Genetics of Personality Consortium (GPC-2) neuroticism GWAS (de Moor *et al.*, 2015) explained 2.75% of the variation in neuroticism in UK Biobank. These results supported reports from twin studies which have shown that around 40–60% (Jang, Livesley and Vemon, 1996; Vukasović and Bratko, 2015) of the variation in the five broad dimensions of personality (neuroticism, extraversion, openness, agreeableness and conscientiousness) is heritable (H^2). Twin based heritability for neuroticism is estimated to be between 56% (women) and 49% (men) (Kendler *et al.*, 2006b; Vukasović and Bratko, 2015), however SNP based heritability assessments suggest that the additive contribution is around 15% in neuroticism (Smith *et al.*, 2016). Due to this substantial missing heritability (Δh^2_{twin}-h^2_{SNP}) it is thought that neuroticism may show a considerable amount of non-additive such as gene-gene interaction or epistatic effects (Jang, Livesley and Vemon, 1996; Loehlin, Neiderhiser and Reiss, 2003; Power and Pluess, 2015; Vukasović and Bratko, 2015).

The importance of epistasis in complex human traits is a long-standing controversy in genetics despite the fact that interactions have been shown to be a critical component of the genomic architecture in animal models (Mackay, 2014; Grice, Liu and Webber, 2015; He *et al.*, 2016). Some researchers claim that human complex traits are the sum of multiple genetic factors and support an additive view of the contribution of individual genes (Hill, Goddard and Visscher, 2008). Others claim that simple additivity does not explain the total amount of variation observed in a trait, so the remaining variation might be due to antagonistic or synergistic interactions between genes (Mackay and Moore, 2014). Standard practice when performing a genome-wide search for genetic association to a trait or disease is to investigate each genomic marker individually and to calculate polygenic scores to assess the relation between an additive component and the outcome. These two approaches ignore the presence of gene-gene or gene-by-environment interactions (Hill, Goddard and Visscher, 2008; Huang and Mackay, 2016). A key point is that using statistical models which operate under certain assumptions (e.g., additivity) may lead to a confirmation bias (Huang and Mackay, 2016). In other words, if an additive model is used then it may confirm the presence of additivity, even if the true genomic architecture is epistatic.

I hypothesise that the effects of single genetic loci, polygenic components and epistasis may all be important in human complex traits – with some traits having more contribution from some components than others, as observed in twin-based broad heritability studies (Loehlin, Neiderhiser and Reiss, 2003; Cesarini and Visscher, 2017). Assessing the relative contribution of the individual components leads to an analysis bottleneck. For example standard methods of testing for epistasis among all possible SNP-SNP combinations is computationally expensive. It also leads to overly conservative multiple testing corrections to balance the large number of combinations tested. A more desirable statistical methodology would allow for simultaneous modelling of single loci, additive and epistatic components. Therefore, the relative contribution of these different components can be assessed while also controlling for the effects of the others. Nicodemus *et al.* (2014) proposed a model that allows for the simultaneous analysis of single markers, additive and epistatic components. This was

a relatively simplistic approach using only SNPs located in and +/- 20kb around genes from the *ZNF804A* pathway and limited to a simple 2-SNP interaction component. Still, two 2-SNP interactions were observed which explained 2-3 times more variation in spatial working memory (SWM) in patients with psychosis than the additive effect. The results were tested for replication in two independent test sets of cases: A) 170 individuals with schizoaffective disorder or schizophrenia and B) 84 individuals with broad psychosis; the model was also applied to controls (*n*=89). In both test sets of cases the R^2 for SWM increased from 1.2% using solely the additive effect to 4.8% when including the two 2-SNP interaction terms observed in the training set. Finally, these interactions term did not explain more variation in the control group.

This project aimed to better understand the complex genetic architecture of neuroticism by extending the work of Nicodemus *et al.* (2014) via our novel statistical methodology, MAICA (Machine-learning for Additive and Interaction Combined Analysis). MAICA is an agnostic methodology assessing single marker, polygenic, gene-gene interactions and environmental components simultaneously. MAICA does not force any of these components into the final model, but rather assesses their contribution using the Least Absolute Shrinkage and Selection Operator (LASSO) model (Tibshirani, 1996). MAICA represents a novel methodology in understanding the genomic architecture of complex human traits in an open-source platform that can be applied to other complex traits.

In this study, we applied MAICA on sets of genetically unrelated individuals from two different cohorts: Generation Scotland (*n*=7,273) and the UK Biobank (*n*=286,800). Using MAICA we tested whether a significant proportion of variation in neuroticism could be explained by assessing all genetic components simultaneously.

4.3 Material and Methods

4.3.1 United Kingdom Biobank (UK-B)

The United Kingdom Biobank is a large population based cohort sampled from 22 locations in the UK (England, Wales and Scotland). UK Biobank was established as a high-powered prospective study to allow for detailed analyses of genetic and non-genetic predictors of diseases and traits more commonly observed in middle and old age (Collins, 2012; Sudlow *et al.*, 2015; Bycroft *et al.*, 2017) by recruiting individuals of the general population of the UK aged 39 to 73. Individuals in this age group who were registered with the National Health Service (NHS) and living no more than 25 miles from an assessment centre were asked to participate. Between 2006 and 2010, over 0.5 million individuals (*n*=502,649) were recruited. Participants completed questionnaires (touch-screen and verbal interview) on their lifestyle (e.g. physical activity), underwent a wide range of physical measurements (e.g. blood pressure and hand grip strength), performed cognitive performance tests and blood, urine and saliva samples were taken. The DNA of UK Biobank participants was genotyped on the Affymetrix UK BiLEVE Axiom array (*n*=49,950) and UK Biobank Axiom array (*n*=438,427) (Bycroft *et al.*, 2017). The combined dataset from both arrays provided 805,426 markers for analysis. Two-thousand-and-eight participants were excluded after genotype quality control (QC) (e.g. identified as outliers for heterozygosity and missingness) and were removed.

A genetically homogenous subgroup of 462,065 individuals using 4-means clustering of the first two genetic principal components was identified representing white British participants (*n*=26,176). Participants overlapping with the PGC MDD (Wray and Sullivan, 2018) and Generation Scotland (Smith *et al.*, 2006; B. H. Smith *et al.*, 2013) datasets (*n*=760) were removed to exclude potential overlap during replication analyses. We removed 131,790 participants with a UK-B reported KING (Manichaikul *et al.*, 2010) kinship coefficient > 0.044. To increase the sample size we create a genomic relationship matrix using Genome-wide Complex Trait Analysis (GCTA) (Yang *et al.*, 2011) on these excluded individuals. A member of each group

of related individuals was added back to the sample by selecting individuals with a genetic relatedness less than 0.025 with any other participant. Leaving a total of 371,437 participants. Of these we removed individuals without a measurement of neuroticism leaving 300,752 participants.

4.3.2 Generation Scotland (GS:SFHS)

Generation Scotland: the Scottish Family Health Study is a, large family-based cohort (Smith *et al.*, 2006; D. J. Smith *et al.*, 2013). Around 24,000 individuals were recruited between 2006 and 2011 from the general population of Scotland. Participants were measured for a wide variety of phenotypes including lifestyle factors and health outcomes. DNA of 20,128 participants was genotyped by a high-density genome-wide bead array (Illumina OmniExpress 700K SNP GWAS and 250K exome chip). Population outlier individuals were removed (Amador *et al.*, 2015). We extracted 7,326 genetically unrelated participants with a valid measurement of neuroticism using GCTA (relatedness < 0.025).

4.3.3 Genetic overlap GS:SFHS and UK-B

In both datasets we removed SNPs and participants with a missingness of >1%, removed SNPs with a minor allele frequency <1% and removed SNPs surpassing the Hardy-Weinberg exact test p-value threshold of 1×10^{-6} (n SNPs removed: GS = 4,361 and UKB = 17,907). This left 7,273 participants in Generation Scotland and 286,800 participants in UK Biobank (Figure 4.1). We selected SNPs common to both UK Biobank and Generation Scotland (n = 49,962). Using PLINK, we performed linkage disequilibrium (LD) pruning (window size = 50kb, step size = 5kb and r^2 threshold = 0.25) leaving 39,406 SNPs in Generation Scotland and 39,515 SNPs in UK Biobank in linkage equilibrium. Of these we removed again the non-overlapping SNPs leaving 38,779 SNPs in both datasets (Figure 4.1). As the statistical methodology is unable to handle missing data, for both datasets missing genotypes were filled by means of median imputation.

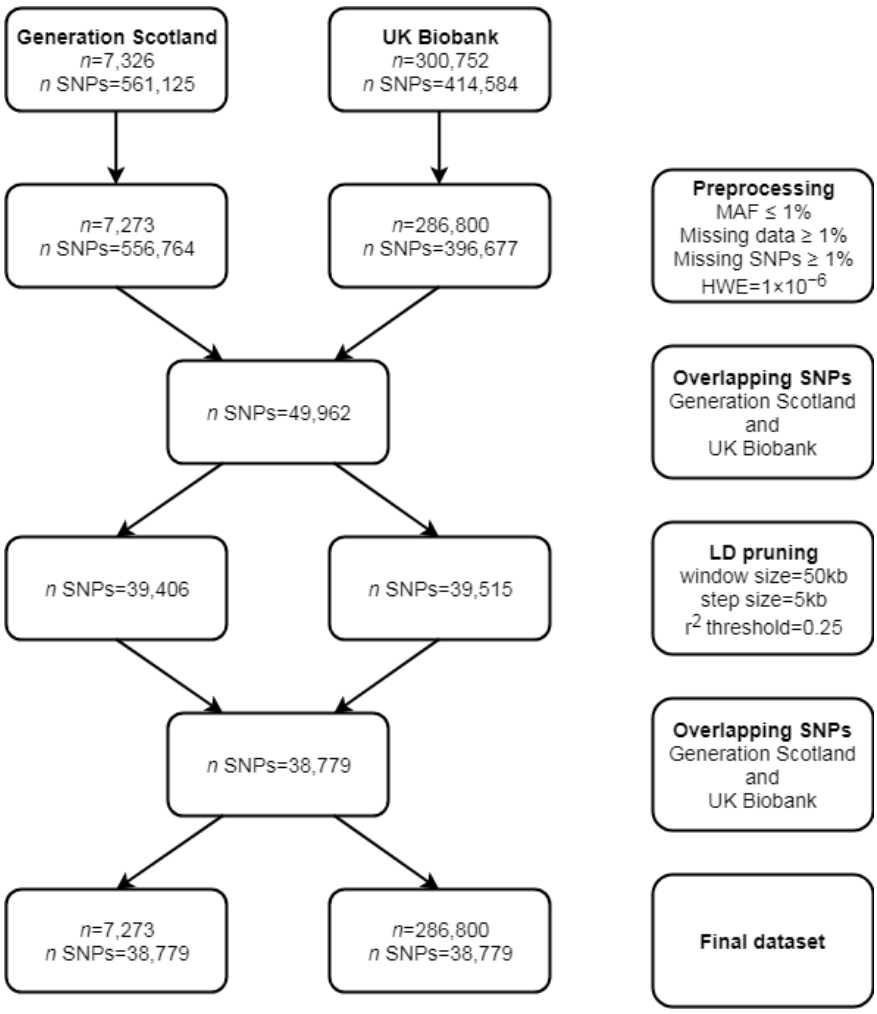


Figure 4.1: Step-by-step representation of the filtering process to ascertain the overlapping SNPs between UK Biobank and Generation Scotland

4.3.4 Neuroticism

The personality trait neuroticism was measured in Generation Scotland and UK-Biobank as the total score of the 12-item Eysenck Personality Questionnaire-Revised (EPQ-R) Short Form (Supplementary Table 4.1) (Eysenck, Eysenck and Barrett, 1985). Neuroticism scores were available for 21,387 participants in Generation Scotland and 401,663 participants in UK Biobank. Scores from the EPQ-R show high internal consistency and overall validity (Matthews, Deary and Whiteman, 2009; Okbay *et al.*, 2016).

4.3.5 MAICA: *Machine-learning for Additive and Interaction Combined Analysis*

I developed a novel methodology incorporated in an R Statistical Computing Environment (R Core Team, 2017) package called MAICA. MAICA separately calculates two genetic components (polygenic effects and gene-gene interactions) and combines this with the third genetic component (genotype data; single SNP effect); these are three out of the four components (the fourth being dominance effects) of broad sense heritability (H^2) (Kempthrone, 1957). Each component's relative contribution can be assessed directly. The genomic predictors are combined, and phenotypic/environmental measurements may be included. LASSO (Least Absolute Shrinkage and Selection Operator), a penalised regression machine learning method, is used to assess the high dimensional data to detect informative components and to form a final model (Tibshirani, 1996). See Figure 4.2.

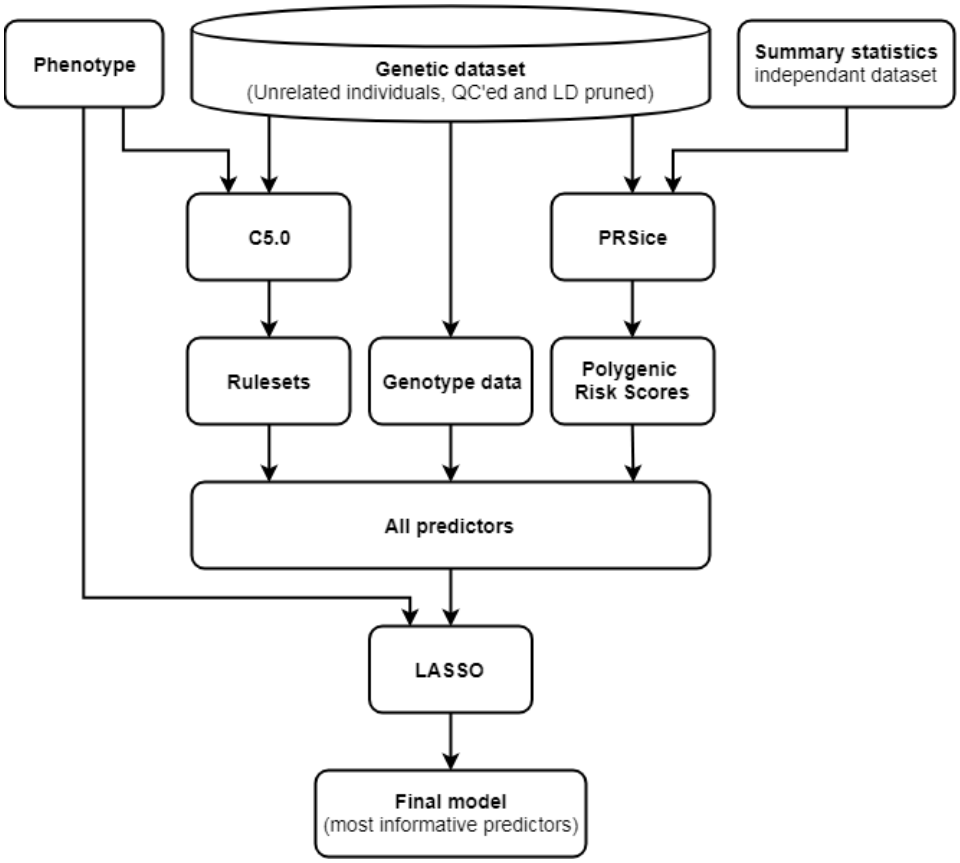


Figure 4.2: Visual representation of MAICA.

4.3.5.1 Single genes

The single gene component used was the standard genotype data, where the count of the number of minor alleles (0, 1, 2) is provided for every SNP.

4.3.5.2 Polygenic effect component

PRSice 2.0 (Euesden, Lewis and O'Reilly, 2015) was incorporated into MAICA to model genetic additivity by means of Polygenic Scores (PRS). PRS are the sum of trait-associated alleles across many loci weighted by effect sizes estimated from an independent Genome-Wide Association Study (GWAS). PRS are calculated over SNPs at a user-specified number of p-value thresholds to represent the polygenic effect. We used the same neuroticism meta-analysis based GWAS summary statistics (*n* individuals=63,000) (de Moor *et al.*, 2015) used by Luciano *et al.* (2018) as the independent dataset. In total, 37,908 SNPs overlapped between our datasets and the independent dataset and those were used to calculate the PRS in eight ranges (P<0.001; 0.01; 0.05; 0.1; 0.2; 0.3; 0.5 and 1).

4.3.5.3 Gene-gene interaction component

MAICA utilises C5.0, a non-parametric algorithm that builds decision trees to model gene-gene interactions (Quinlan, 1992; Salzberg, 1994; Kuhn and Johnson, 2013; Zhang *et al.*, 2015). C5.0 constructs a large decision tree where every top to bottom path is called a "ruleset", which can be interpreted as a sequence of genetic interactions. This tree is pruned to remove all rulesets that are not informative. The remaining rulesets can be used as predictors in a regression model. Previous work chapter 3 (Meijsen *et al.*, 2018) has shown that C5.0 is capable of accurately detecting interacting SNPs.

4.3.5.4 Least Absolute Shrinkage and Selection Operator (LASSO)

LASSO is a penalised regression machine learning method used to simultaneously assess the contribution of all genetic components in high dimensional data. In short,

LASSO fits a model between the outcome and the predictors much like ordinary least squares. However, in contrast to ordinary least squares, LASSO limits the sum of absolute values of coefficients to not be larger than a constant value. To achieve this LASSO shrinks the coefficients of the non-informative predictors. When fitting the model, the coefficients of predictors that do not improve the model are shrunken by means of the penalty parameter λ. The λ corresponding with the lowest mean square error was selected by means of a 10-fold nested cross validation analysis. For a more in-depth outline of LASSO see *section 1.7.1*.

4.3.5.5 Phenotypic/environmental measurements

In line with Luciano *et al.* (2018) a linear regression analysis was performed for neuroticism scores controlling for sex, age, assessment centre, genotype batch, array and 40 genetic principal components (PC). The residuals were extracted to act as an adjusted neuroticism measurement (Zhao *et al.*, 2012). We used the residuals of Generation Scotland using only sex and age due to its homogeneous population and sampling method. The difference in R^2 between the neuroticism model adjusted for sex and age and the model adjusting for the first 15 PCs is 0.25% therefore the "increase" from PC6 and PC7 is negligible and due to scaling (Supplementary Figure 4.1A-D). We applied MAICA on a training and testing set, or discovery and replication set, study design as are routinely carried out in a machine learning context. First, we applied MAICA on the Generation Scotland dataset (*n*=7,273) as the training set and used the full UK Biobank dataset (*n*=286,800) as an independent test set; second, we applied MAICA on a randomly-sampled 50% UK Biobank training set (*n*=143,400) and used the remaining 50% of UK Biobank participants (*n*=143,400) for the independent test set.

4.4 Results

4.4.1 Applying MAICA on GS with replication in UK-B

Applying C5.0 (incorporated in MAICA) on the residualised neuroticism scores in Generation Scotland (*n*=7,273), yielded 17 rulesets containing in total 16 SNPs associated with the trait (Supplementary Table S4.2). We assigned all individuals to rulesets based on their genotype data resulting in 17 novel predictors. These were combined with the genotype data and polygenic scores forming a dataset of 38,804 predictors. We performed a nested 10-fold cross-validation in LASSO to ascertain the λ setting corresponding with the lowest mean square error. We subsequently reapplied LASSO using the optimal λ setting (λ = 0.10) resulting in 162 predictors with a non-zero coefficient (Supplementary Table 4.3). Of the 162 predictors selected by LASSO, the majority were single SNPs (*n*=151) while the remaining 11 predictors were rulesets. No PRS were selected for inclusion in the model. The predictors were able to account for 6.6% (R^2 =0.066; *p*-value < 1.0×10^{-16}) of variation in neuroticism scores in the Generation Scotland training set. Using the predict option in MAICA we used the selected features to predict the neuroticism scores in UK Biobank. However the predicted outcomes did not correlate with the observed outcome (r = 7.0×10^{-04}, *p*-value = 0.71), suggesting that the model created on the training data is likely overfitting the data.

4.4.2 Applying MAICA on UK Biobank

We randomly divided the UK Biobank into independent training and test sets. Using the training set (*n*=143,400) C5.0 returned 100 rulesets containing 99 SNPs associated with neuroticism. Combined with the polygenic scores and genotype data led to a dataset of 38,887 predictors. LASSO returned no predictors with non-zero coefficients after a 10-fold cross validation, so no further analysis was performed.

4.5 Discussion

By using Machine-learning for Additive and Interaction Combined Analysis (MAICA) we were able to simultaneously assess the contribution of three out of the four components making up broad sense heritability (single marker, polygenic and gene-gene interactions). Our method applies a penalised regression machine learning technique that selects only informative components to be in the final model. Using MAICA, we were able to explain a significant amount of variation in neuroticism in Generation Scotland, but these results were likely inflated due to overfitting by using the same training data to derive both the rulesets and the final model.

We were not able to replicate GS results in UK-B as predicted neuroticism scores using the 162 predictors observed in Generation Scotland did not correlate with the observed neuroticism scores ($r = 7.0x10^{-04}$, p-value = 0.71). The most plausible explanation for this difference is - due to GS being around 40 times smaller than UK-B – that MAICA overfit the data which is not uncommon in smaller datasets. Other reasons might be that due to the size difference the model will explain a significant proportion of variation in a specific subset of individuals of UK-B as in general smaller datasets show less variation than large datasets. Therefore it is possible that the model created using GS might be accurate but explains little variation due to the larger amount of variation in UK-B. It has to be noted that even though the training set - Generation Scotland- is significantly smaller than the test set it is a replication analysis between two independent cohorts which is a point of strength compared to a replication analysis within the same cohort. MAICA was not able to create a model using the UK Biobank training set and therefore no replication on the test set was applied. One of the more notable disadvantages of LASSO is bound by the number of samples, therefore having a dataset with more predictors than samples is problematic (Zou and Hastie, 2005), which is the case in Generation Scotland only. However, LASSO did not select any predictor in the UK Biobank train- and test set design therefore this explanation seems unlikely.

For both Generation Scotland and UK Biobank C5.0 (the epistatic model used by MAICA) detected several rulesets (GS = 16 and UK-B = 100). In previous work we

have shown that C5.0 has a conservative type I error and high power. Reasons why MAICA did not include any rulesets in UK Biobank compared to GS might be again that MAICA overfit to the data. We used SNPs that overlapped between Generation Scotland and UK Biobank. Therefore, we might have removed SNPs that were associated with neuroticism during the LD pruning process. To investigate this possibility, we assessed the overlap between our UK Biobank dataset and the 116 neuroticism GWAS hits previously observed in UK Biobank. Of these, we found that only two overlapped, the remaining 114 were not included in our dataset. Further, we compared the amount of variation explained by our PRS analysis of UK Biobank data and to that of a previous study using UK Biobank. Even though we used the same dataset and the same independent summary statistics, our PRS's explain at most 0.1% (p-value range 0-1) of variation compared to around 2.75% mentioned in Luciano *et al.* (2018). This strengthens the assumption that we have potentially removed most of the SNPs that are strongly associated with neuroticism.

4.6 Conclusion

In conclusion, we have shown that MAICA explained a small amount of variation in Generation Scotland which is most likely inflated due to overfitting and we were not able to predict neuroticism scores in UK Biobank using the Generation Scotland model. The strength of the Generation Scotland - UK Biobank replication study is that both are independent cohorts even though Generation Scotland is many times smaller than UK Biobank. MAICA was not able to create a model using the UK Biobank training/test set study design. Future work will be to re-examine this model using imputed data from Generation Scotland so that the overlapping SNPs between datasets will not remove associated regions which was observed in this analysis. It is possible that neuroticism does not have a strong epistatic component as expected however we will apply MAICA to other phenotypes to see if models differ.

5 Discussion

5.1 Summary of aims of thesis

The aims of this thesis were two-fold. The first was to examine whether cognitive performance differs between Major Depressive Disorder (MDD) cases and controls and whether recurrent depression cases differ from single-episode cases. We examined genomic associations using conventional methods - e.g. Genome-Wide Association Studies (GWAS), Genome-Wide Environmental Interaction Studies (GWEIS) and Polygenic Scores (PRS) - with cognitive differences as a depression endophenotype strategy. The aim of the second part of this thesis was to develop statistical methodology that allows simultaneous modelling of three out of the four components making up broad sense heritability (single loci, polygenic and gene-gene interactions). These can be assessed by a machine learning algorithm to detect informative components. Specifically, the application of machine learning to genetics might help us to gain a better understanding of the genetic mechanisms and aetiologies and explain more variation in complex human traits.

5.2 Summary of findings

5.2.1 Chapter 2

The work presented in Chapter 2 outlined a standard approach into investigating the phenotypic and genetic differences of a trait of interest. We have shown in this work that applying some standard approaches are effective in detecting genetic associations and explained small proportions of phenotypic variation (e.g. PRS). Others standard methods were not (e.g. GWAS and GWEIS), all of which are important in understanding the genetic underpinnings of a complex human trait such as MDD or cognitive ability. However, it is crucial to understand - and this has been addressed on numerous occasions in this thesis - that by focussing solely on GWAS (single loci) and PRS (polygenic) in combination with models including the additive

effects of environmental factors will exclude the possible contribution of non-additive effects (interactions). We confirmed in Chapter 2 that cognitive performance in MDD cases did indeed deviate from healthy controls and also deviated between MDD subtypes (single episode and recurrent episode MDD). We observed that participants with recurrent MDD had significantly slower processing speeds compared to controls and participants with single episode MDD. Moreover, MDD cases, particularly recurrent MDD cases, scored significantly higher on the vocabulary test than controls. This had been previously reported in the larger UK Biobank cohort; however, the UK Biobank cohort relied on self-reported mental health, whereas the Generation Scotland cohort provided clinically diagnosed MDD. The differences observed might be real, however differential recall and higher health literacy. Other explanations for the difference in vocabulary scores. No genome-wide associations or interactions were observed to explain these differences. Using a large processing speed meta-analysis (Ibrahim-Verbaas *et al.*, 2016) we were able to account for up to 1% of phenotypic variation in processing speed using polygenic scores in Generation Scotland. The material presented in Chapter 2 has been peer reviewed and has been published in *Translational Psychiatry*.

5.2.2 Chapter 3

Performing an exhaustive analysis (e.g. analyse all possible combinations) to detect genetic interactions is not inherently desirable. When assessing all possible combinations, the amount of computational power and time needed is high. The large number of variables in most genetic studies – leading to an even larger number of combinations of these variables - requires a conservative statistical threshold to control for multiple testing. We hypothesised that applying non-parametric tree based methods is one potential solution to these issues, as they explicitly model interactions and do not need predefined settings, such as the number of variables to consider in an interaction. In this chapter we performed an extensive simulation study to assess the capability of two non-parametric tree based methods to detect simulated interacting SNPs, then applied the methods to educational attainment in Generation Scotland: the Scottish Family Health Study. Assessing the power of both methods by

simulating interactions of different sizes and interaction strength levels is key, as it provides a benchmark for assessing a method's power and type I error. Non-parametric methods use a greedy search of all variables available in the dataset. Therefore, they are still performing many tests, but not an exhaustive search. This chapter was a proof-of-principle study to show the capability of these methods to detect interactions, assigning individuals to rulesets or logic trees as predictors and evaluating the power and type I error in a high dimensional dataset containing other genetic components (e.g. single SNPs and PRS). In Chapter 3, a total of six epistasis models were created varying in size and association strength level, each containing 500 replicates. By applying this study design we have shown that C5.0 and logic regression were both capable in detecting simulated genetic interactions using a wide range of association strength levels combined with a strong polygenic component. Applying LD pruning on the dataset prior to analysis helped to improve the power and reduced the type I error for both methods. Finally, we suggested using C5.0 over logic regression as C5.0 showed a more conservative type I error and higher power compared to logic regression. These results were supported when applying both methods on years of education as a proxy for educational attainment in Generation Scotland. C5.0 was able to detect numerous interacting SNPs including SNPs located in genes associated with learning, reading and neurodevelopment, while the model created by logic regression did not pass the randomisation test (α=0.05). The material presented in Chapter 3 has been peer reviewed and has been published in *Bioinformatics*.

5.2.3 Chapter 4

The work presented in Chapter 4 combines all methodologies addressed in Chapters 2 and 3. As mentioned throughout this thesis, epistasis is widely known to be involved in complex traits in animals, yet the debate regarding its importance in humans is ongoing. Developing an agnostic methodology called *MAICA* (Machine-learning for Additive and Interaction Combined Analysis) that simultaneously models the components used throughout this thesis and applying machine learning to build the final model will provide software to further discussions in this debate.

MAICA represents a novel idea and methodology in explaining the contribution of multiple genetic components to the trait as an extension to the method previously proposed by Nicodemus *et al.* (2014). In chapter 4 we applied MAICA to neuroticism among unrelated individuals in both Generation Scotland and UK Biobank, using the overlapping SNPs between both studies after quality control and linkage disequilibrium pruning. MAICA returned 162 predictors (151 single SNPs and 11 interactions) using Generation Scotland. These components explained 6.62% of variation in neuroticism scores in Generation Scotland. The model did not predict neuroticism scores in UK Biobank ($r = 7.0x10^{-04}$, p-value = 0.706). This large difference is most likely due to MAICA overfitting the data in Generation Scotland and the difference in dataset size between Generation Scotland and UK Biobank. UK Biobank is around 40 times larger than Generation Scotland this in turn increases the amount of variation. The model created in Generation Scotland might explain a larger amount of variation in a subgroup of individuals in UK Biobank. Applying MAICA on the training set of the UK Biobank set (*n*=143,400) we were not able to detect any component associated with neuroticism, therefore we did not perform a replication analysis on the test set. It is possible that the genetic signal was removed after LD pruning on only the overlapping SNPs between Generation Scotland and UK Biobank.

5.3 Strengths

The work presented in this thesis addresses two separate - but linked - challenges in the field. The first challenge is more ideological, as researchers have long debated the existence and importance of epistasis in humans. The debate regarding epistasis revolves mainly around two opposing views, with one side claiming that complex human traits are simply the sum of multiple independently-acting genetic factors while others claim that this principle does not explain the full amount of genetic variance, leaving room for deviations from this additivity due to non-independently acting genetic factors. Utilising methods that solely test for one of the views may result in a confirmation bias, as other possibilities are not assessed. In this thesis we have created a methodology that tries to fill the void by finding a middle ground. Combining single

loci, polygenic scores and gene-gene interactions in one design matrix and applying machine learning to detect informative predictors allows for users to empirically test for all components. In doing so, we hoped to fill the gap of missing heritability. The first challenge was addressed in chapter 3 as we showed that C5.0 was able to detect the simulated epistasis created with high accuracy and conservative type I error. The second challenge is statistical; many attempts to address epistasis have suffered from the curse of dimensionality where potentially true associations were deemed non-significant due to the stringent statistical penalty required after testing numerous combinations of genetic components. It has to be noted that in most situations MAICA will overfit the model to the data, and, like in most machine learning study designs, will always require an independent test set to evaluate performance. This was observed in the smaller dataset Generation Scotland where MAICA explained 6.62% variance, when applied to the larger independent dataset UK Biobank, the model did not predict neuroticism outcomes. Arguments can be made that this is a point of strength as we applied MAICA on different independent cohorts rather than within cohort. However, this was only the first application of MAICA, and so future work will apply the method to other phenotypes and data sets to assess performance.

5.4 Caveats

It is important to highlight potential caveats to this thesis. We found some evidence for the existence of epistasis but this is not unambiguous. We simulated a polygenic and interactions separately to combine the signals. By doing this we assumed that the components acted independently; however, it would be possible that single markers or polygenic scores might interact with each other, this was not modelled in chapter 3. There might even be a mixture between additive and epistatic components which make up a polygenic signal (Webber, 2017). This is something that should be considered in future work. In this thesis we have solely used single loci, polygenic scores and epistatic components, however we have excluded potential other components e.g. gene-by-environment interaction, mitochondrial DNA (mtDNA) interactions and epigenetics which can be included in follow up studies. In Chapter 4 we apply MAICA on the overlapping SNPs between Generation Scotland and UK Biobank after Linkage

Disequilibrium (LD) pruning, however this might have removed the true signal and therefore might explain why no components were selected in the UK Biobank train and test set study design. Also training MAICA on a "small" dataset such as Generation Scotland and testing on a large dataset such as UK Biobank might explain why the amount of variation explained in the train and test set differs drastically as larger datasets in general show more variation than smaller datasets. LASSO does show issues when the number of predictors outnumbers the amount of samples, which was the case in the Generation Scotland training set but not in the UK Biobank training set analysis.

5.5 Future work

This section will outline potential follow-up studies extending the work described in this thesis. The most important topics to address have been mentioned in the caveat section. Performing an extensive simulation study of multiple (small) epistatic models to assess whether the methods are capable of detecting an additive-epistasis component would be possible. This may be biologically plausible however no study has been published reporting this type of association in any living organism, making it an exploratory study. We observed that LD pruning has a positive impact on the percentage of accurately detected simulated interactions. The downside of LD pruning is that a large proportion of data which may be biologically relevant is removed. We investigated the possibility to prevent this loss of data due to LD pruning and discovered an existing methodology called "LD sub-setting" (Walters, Laurin and Lubke, 2012). This method assesses the LD structure of a genetic dataset and splits SNPs into separate SNP subsets in Linkage Equilibrium (LE) that can be analysed individually. During this work we encountered computational limitations related to working with large genetic datasets (e.g. UK Biobank; n=500,000). The software for MAICA was written in the R Statistical Computing Environment. For it to be used with very large data sets, it will need to be re-coded in a computationally efficient language such as C++. Previously, data have typically been analysed in a single database by one computer; essentially, analyses were sequential. With the dimensionality of data increasing exponentially in large-scale biobanking efforts, this

has created a choke point as all data needs to be analysed by the same machine for which historically the only solution is the increase the memory of the machines. To address this problem, Google developed an algorithm called MapReduce that allows for large data to be divided into smaller subsets and mapped to many computers, creating parallel analyses. When the analysis is complete the results are mapped back together to produce the final result. To optimise MAICA for use with ever-growing data, integrating with MapReduce is essential.

MAICA addresses both the ideological and statistical debate commonly used against testing for epistasis. MAICA is a valuable assets to assess the different components contributing to variation in complex human traits.

6 References

Amador, C. *et al.* (2015) 'Recent genomic heritage in Scotland', *BMC Genomics*, 16(1). doi: 10.1186/s12864-015-1605-2.

American Psychiatric Association (1994) 'American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition', *American Psychiatric Association*.

Anderson, R. J., Freedland, K. E. and Lustman, P. J. (2001) 'The prevalence of comorbid depression in adults with diabetes: A meta-analysis', *Diabetes Care*, 24(6), pp. 1069–1078. doi: 10.2337/diacare.24.6.1069.

Baddeley, A. (2007) *Working Memory, Thought, and Action*. Oxford University Press. doi: 10.1093/acprof:oso/9780198528012.001.0001.

Barry, D., Bates, M. E. and Labouvie, E. (2008) 'FAS and CFL Forms of Verbal Fluency Differ in Difficulty: A Meta-analytic Study', *Applied Neuropsychology*. NIH Public Access, 15(2), pp. 97–106. doi: 10.1080/09084280802083863.

Bateson, W. and Mendel, G. (1909) 'Mendel's Principles of heredity, by W. Bateson'. Cambridge : University Press, pp. 1–450. doi: 10.5962/bhl.title.44575.

Bosker, F. J. *et al.* (2011) 'Poor replication of candidate genes for major depressive disorder using genome-wide association data', *Molecular Psychiatry*, 16(5), pp. 516–532. doi: 10.1038/mp.2010.38.

Bouchard, T. and McGue, M. (1981) 'Familial studies of intelligence: a review', *Science*, 212(4498), pp. 1055–1059. doi: 10.1126/science.7195071.

Brockmann, G. A. *et al.* (2000) 'Single QTL effects, epistasis, and pleiotropy account for two-thirds of the phenotypic F2 variance of growth and obesity in DU6i x DBA/2 mice', *Genome Research*. Cold Spring Harbor Laboratory Press, 10(12), pp. 1941–1957. doi: 10.1101/gr.GR1499R.

Bycroft, C. *et al.* (2017) 'Genome-wide genetic data on ~500,000 UK Biobank participants', *bioRxiv*. Cold Spring Harbor Laboratory, p. 166298. doi:

10.1101/166298.

Cai, N. *et al.* (2015) ‘Sparse whole-genome sequencing identifies two loci for major depressive disorder’, *Nature*. Nature Research, 523(7562), pp. 588–591. doi: 10.1038/nature14659.

Carroll, J. B. (1993) *Human Cognitive Abilities: A Survey of Factor-Analytic Studies, Educational Researcher*. doi: 10.1017/CBO9780511486371.

Cesarini, D. and Visscher, P. M. (2017) ‘Genetics and educational attainment’, *npj Science of Learning*, 2(1), p. 4. doi: 10.1038/s41539-017-0005-6.

Chan, R. C. K. *et al.* (2008) ‘Assessment of executive functions: Review of instruments and identification of critical issues’, *Archives of Clinical Neuropsychology*, 23(2), pp. 201–216. doi: 10.1016/j.acn.2007.08.010.

Chan, S. W. Y., Goodwin, G. M. and Harmer, C. J. (2007) ‘Highly neurotic never-depressed students have negative biases in information processing’, *Psychological Medicine*, 37(9), pp. 1281–1291. doi: 10.1017/S0033291707000669.

Cheng, Y. *et al.* (2011) ‘Mapping genetic loci that interact with myostatin to affect growth traits’, *Heredity*. Nature Publishing Group, 107(6), pp. 565–573. doi: 10.1038/hdy.2011.45.

Collins, R. (2012) ‘What makes UK Biobank special?’, *The Lancet*. Elsevier, pp. 1173–1174. doi: 10.1016/S0140-6736(12)60404-8.

Conley, J. J. (1985) ‘Longitudinal Stability of Personality Traits. A Multitrait-Multimethod-Multioccasion Analysis’, *Journal of Personality and Social Psychology*, 49(5), pp. 1266–1282. doi: 10.1037/0022-3514.49.5.1266.

Cosentino, S., Manly, J. and Mungas, D. (2007) ‘Do reading tests measure the same construct in multiethnic and multilingual older persons?’, *Journal of the International Neuropsychological Society : JINS*. NIH Public Access, 13(2), pp. 228–36. doi: 10.1017/S1355617707070257.

Cuijpers, P. *et al.* (2010) ‘Economic costs of neuroticism: a population-based study.’,

Archives of General Psychiatry, 67(10), pp. 1086–1093. doi: 10.1001/archgenpsychiatry.2010.130.

Cullen, B. *et al.* (2015) 'Cognitive function and lifetime features of depression and bipolar disorder in a large population sample: Cross-sectional study of 143,828 UK Biobank participants', *European Psychiatry*, 30(8), pp. 950–958. doi: 10.1016/j.eurpsy.2015.08.006.

Davies, G. *et al.* (2015) 'Genetic contributions to variation in general cognitive function: a meta-analysis of genome-wide association studies in the CHARGE consortium (N=53949).', *Molecular psychiatry*. Nature Publishing Group, 20(2), pp. 183–92. doi: 10.1038/mp.2014.188.

Davies, G. *et al.* (2016) 'Genome-wide association study of cognitive functions and educational attainment in UK Biobank (N= 112 151)', *Molecular Psychiatry*. Nature Publishing Group, 21(August 2015), pp. 1–10. doi: 10.1038/mp.2016.45.

Deary, I. J. *et al.* (2002) 'Cognitive change and the APOE epsilon 4 allele.', *Nature*, 418(6901), p. 932. doi: 10.1038/418932a.

Deary, I. J., Harris, S. E., *et al.* (2005) 'KLOTHO genotype and cognitive ability in childhood and old age in the same individuals', *Neuroscience Letters*, 378(1), pp. 22–27. doi: 10.1016/j.neulet.2004.12.005.

Deary, I. J., Hamilton, G., *et al.* (2005) 'Nicastrin gene polymorphisms, cognitive ability level and cognitive ageing', *Neuroscience Letters*, 373(2), pp. 110–114. doi: 10.1016/j.neulet.2004.09.073.

Deary, I. J. (2012) 'Intelligence', *Annual Review of Psychology*, 63(1), pp. 453–482. doi: 10.1146/annurev-psych-120710-100353.

Deary, I. J. (2014) 'The Stability of Intelligence From Childhood to Old Age', *Current Directions in Psychological Science*. SAGE PublicationsSage CA: Los Angeles, CA, 23(4), pp. 239–245. doi: 10.1177/0963721414536905.

Deary, I. J., Johnson, W. and Houlihan, L. M. (2009) 'Genetic foundations of human intelligence', *Human Genetics*, pp. 215–232. doi: 10.1007/s00439-009-0655-4.

Deary, I. J., Penke, L. and Johnson, W. (2010) ‘The neuroscience of human intelligence differences’, *Nature Reviews Neuroscience*. Nature Publishing Group, 11(3), p. 201. doi: 10.1038/nrn2793.

Debette, S. *et al.* (2015) ‘Genome-wide studies of verbal declarative memory in nondemented older people: The Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium’, *Biological Psychiatry*. NIH Public Access, 77(8), pp. 749–763. doi: 10.1016/j.biopsych.2014.08.027.

Delgado, V. B., Kapczinski, F. and Chaves, M. L. F. (2012) ‘Memory mood congruency phenomenon in bipolar I disorder and major depression disorder patients’, *Brazilian Journal of Medical and Biological Research*. Associação Brasileira de Divulgação Científica, 45(9), pp. 856–861. doi: 10.1590/S0100-879X2012007500098.

Demirkan, A. *et al.* (2011) ‘Genetic risk profiles for depression and anxiety in adult and elderly cohorts’, *Molecular Psychiatry*, 16(7), pp. 773–783. doi: 10.1038/mp.2010.65.

Dempster, F. N. (1978) ‘Memory span and short-term memory capacity: A developmental study’, *Journal of Experimental Child Psychology*, 26(3), pp. 419–431. doi: 10.1016/0022-0965(78)90122-4.

Diamond, A. (2014) ‘Executive Functions’, *Annual review of clinical psychologyPsychol.* NIH Public Access, 64, pp. 135–168. doi: 10.1146/annurev-psych-113011-143750.Executive.

Eaton, W. W. *et al.* (2008) ‘Population-Based Study of First Onset and Chronicity in Major Depressive Disorder’, *Archives of General Psychiatry*. NIH Public Access, 65(5), p. 513. doi: 10.1001/archpsyc.65.5.513.

Euesden, J., Lewis, C. M. and O’Reilly, P. F. (2015) ‘PRSice: Polygenic Risk Score software’, *Bioinformatics*, 31(9), pp. 1466–1468. doi: 10.1093/bioinformatics/btu848.

Eysenck, S. B. G., Eysenck, H. J. and Barrett, P. (1985) ‘A revised version of the

psychoticism scale', *Personality and Individual Differences*, 6(1), pp. 21–29. doi: 10.1016/0191-8869(85)90026-1.

Falconer, D. S. and Mackay, T. F. . (1996) 'Introduction to quantitative genetics.' Longman, p. 463. doi: 10.1002/bimj.19620040211.

Farmer, A. *et al.* (2008) 'Medical disorders in people with recurrent depression', *British Journal of Psychiatry*, 192(5), pp. 351–355. doi: 10.1192/bjp.bp.107.038380.

Fernandez-Pujals, A. M. *et al.* (2015) 'Epidemiology and heritability of major depressive disorder, stratified by age of onset, sex, and illness course in generation Scotland: Scottish family health study (GS: SFHS)', *PLoS ONE*. Edited by K. Ebmeier, 10(11), p. e0142197. doi: 10.1371/journal.pone.0142197.

Fisher, R. A. (1919) 'XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance.', *Transactions of the Royal Society of Edinburgh*, 52(2), pp. 399–433. doi: 10.1017/S0080456800012163.

Flint, J. and Kendler, K. S. (2014) 'The Genetics of Major Depression', *Neuron*. Elsevier, pp. 484–503. doi: 10.1016/j.neuron.2014.01.027.

Gershon, E. S. (1982) 'A Family Study of Schizoaffective, Bipolar I, Bipolar II, Unipolar, and Normal Control Probands', *Archives of General Psychiatry*, 39(10), p. 1157. doi: 10.1001/archpsyc.1982.04290100031006.

Gibson, G. (2010) 'Hints of hidden heritability in GWAS', *Nature Genetics*. Nature Research, 42(7), pp. 558–560. doi: 10.1038/ng0710-558.

Golden, S. H. *et al.* (2008) 'Examining a bidirectional association between depressive symptoms and diabetes', *JAMA*, 299(23), pp. 2751–9. doi: 10.1001/jama.299.23.2751.

Gottfredson, L. S. (1997) 'Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography', *Intelligence*, 24(1), pp. 13–23. doi: 10.1016/S0160-2896(97)90011-8.

Greenberg, P. E. *et al.* (2003) 'The economic burden of depression in the United

States: How did it change between 1990 and 2000?', *Journal of Clinical Psychiatry*, 64(12), pp. 1465–1475. doi: 10.4088/JCP.v64n1211.

Grice, S. J., Liu, J. L. and Webber, C. (2015) 'Synergistic Interactions between Drosophila Orthologues of Genes Spanned by De Novo Human CNVs Support Multiple-Hit Models of Autism', *PLoS Genetics*. Edited by S. Shifman, 11(3), p. e1004998. doi: 10.1371/journal.pgen.1004998.

Halvorsen, M. *et al.* (2012) 'Cognitive function in unipolar major depression: A comparison of currently depressed, previously depressed, and never depressed individuals', *Journal of Clinical and Experimental Neuropsychology*, 34(7), pp. 782–790. doi: 10.1080/13803395.2012.683853.

Hardeveld, F. *et al.* (2013) 'Recurrence of major depressive disorder and its predictors in the general population: Results from the Netherlands Mental Health Survey and Incidence Study (NEMESIS)', *Psychological Medicine*, 43(1), pp. 39–48. doi: 10.1017/S0033291712002395.

Haworth, C. M. A. *et al.* (2010) 'The heritability of general cognitive ability increases linearly from childhood to young adulthood', *Molecular Psychiatry*, 15(11), pp. 1112–1120. doi: 10.1038/mp.2009.55.

He, X. *et al.* (2016) 'Epistatic partners of neurogenic genes modulate Drosophila olfactory behavior', *Genes, Brain and Behavior*, 15(2), pp. 280–290. doi: 10.1111/gbb.12279.

Hill, M. J. *et al.* (2012) 'Knockdown of the psychosis susceptibility gene ZNF804A alters expression of genes involved in cell adhesion', *Human Molecular Genetics*, 21(5), pp. 1018–1024. doi: 10.1093/hmg/ddr532.

Hill, W. G., Goddard, M. E. and Visscher, P. M. (2008) 'Data and theory point to mainly additive genetic variance for complex traits', *PLoS Genetics*, 4(2). doi: 10.1371/journal.pgen.1000008.

Howard, D. M. *et al.* (2017) 'Genome-wide association study of depression phenotypes in UK Biobank (n = 322,580) identifies the enrichment of variants in

excitatory synaptic pathways', *bioRxiv*. doi: 10.1101/168732.

Huang, W. *et al.* (2012) 'Epistasis dominates the genetic architecture of Drosophila quantitative traits', *Proceedings of the National Academy of Sciences*, 109(39), pp. 15553–15559. doi: 10.1073/pnas.1213423109.

Huang, W. and Mackay, T. F. C. (2016) 'The Genetic Architecture of Quantitative Traits Cannot Be Inferred from Variance Component Analysis', *PLoS Genetics*. Edited by X. Zhu, 12(11), p. e1006421. doi: 10.1371/journal.pgen.1006421.

Ibrahim-Verbaas, C. A. *et al.* (2016) 'GWAS for executive function and processing speed suggests involvement of the CADM2 gene', *Molecular Psychiatry*, 21(2), pp. 189–197. doi: 10.1038/mp.2015.37.

Jaeggi, S. M. *et al.* (2008) 'Improving fluid intelligence with training on working memory.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 105(19), pp. 6829–6833. doi: 10.1073/pnas.0801268105.

Jang, K. L., Livesley, W. J. and Vemon, P. A. (1996) 'Heritability of the Big Five Personality Dimensions and Their Facets: A Twin Study', *Journal of Personality*, 64(3), pp. 577–592. doi: 10.1111/j.1467-6494.1996.tb00522.x.

Kane, M. J. *et al.* (2007) 'Working Memory , Attention Control , and the N -Back Task: A Question of Construct Validity', *Journal of Experimental Psychology: Learning, Memory and Cognition*, 33(3), pp. 615–622. doi: 10.1037/0278-7393.33.3.615.

Kempthrone (1957) 'An introduction to genetic statistics'.

Kendler, K. S. *et al.* (2006a) 'A Swedish national twin study of lifetime major depression', *American Journal of Psychiatry*, 163, pp. 109–114. doi: 10.1176/appi.ajp.163.1.109.

Kendler, K. S. *et al.* (2006b) 'Personality and Major Depression', *Archives of General Psychiatry*, 63(10), p. 1113. doi: 10.1001/archpsyc.63.10.1113.

Kessler, R. C. *et al.* (1993) ‘Sex and depression in the National Comorbidity Survey I: Lifetime prevalence, chronicity and recurrence’, *Journal of Affective Disorders*, 29(2–3), pp. 85–96. doi: 10.1016/0165-0327(93)90026-G.

Kessler, R. C. and Bromet, E. J. (2013) ‘The Epidemiology of Depression Across Cultures’, *Annual Review of Public Health*. NIH Public Access, 34(1), pp. 119–138. doi: 10.1146/annurev-publhealth-031912-114409.

Knol, M. J. *et al.* (2006) ‘Depression as a risk factor for the onset of type 2 diabetes mellitus. A meta-analysis’, *Diabetologia*, 49(5), pp. 837–845. doi: 10.1007/s00125-006-0159-x.

Kooperberg, C. *et al.* (2001) ‘Sequence analysis using logic regression.’, *Genetic epidemiology*, 21 Suppl 1(Suppl 1), pp. S626-31.

Kruglyak, L. *et al.* (1996) ‘Parametric and nonparametric linkage analysis: a unified multipoint approach.’, *American journal of human genetics*, 58(6), pp. 1347–63. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1915045/pdf/ajhg00019-0253.pdf (Accessed: 8 August 2017).

Kuehner, C. (2003) ‘Gender differences in unipolar depression: an update of epidemiological findings and possible explanations’, *Acta Psychiatrica Scandinavica*, 108(3), pp. 163–174. doi: 10.1034/j.1600-0447.2003.00204.x.

Kuhn, M. and Johnson, K. (2013) *Applied Predictive Modeling*. doi: 10.1007/978-1-4614-6849-3.

Lahey, B. B. (2009) ‘Public Health Significance of Neuroticism’, *American Psychologist*, 64(4), pp. 241–256. doi: 10.1037/a0015309.

Lim, J. *et al.* (2013) ‘Sensitivity of cognitive tests in four cognitive domains in discriminating MDD patients from healthy controls: a meta-analysis’, *International Psychogeriatrics*, 25(9), pp. 1543–1557. doi: 10.1017/S1041610213000689.

Loehlin, J. C., Neiderhiser, J. M. and Reiss, D. (2003) ‘The behavior genetics of personality and the NEAD study’, *Journal of Research in Personality*. Academic

Press, 37(5), pp. 373–387. doi: 10.1016/S0092-6566(03)00012-6.

Lopez, A. D. and Murray, C. C. J. L. (1998) ‘The global burden of disease, 1990–2020’, *Nature Medicine*. Nature Publishing Group, 4(11), pp. 1241–1243. doi: 10.1038/3218.

Lubke, G. H. *et al.* (2012) ‘Estimating the genetic variance of major depressive disorder due to all single nucleotide polymorphisms’, *Biological Psychiatry*, 72(8), pp. 707–709. doi: 10.1016/j.biopsych.2012.03.011.

Luciano, M. *et al.* (2006) ‘Genome-wide scan of IQ finds significant linkage to a quantitative trait locus on 2q’, *Behavior Genetics*, 36(1), pp. 45–55. doi: 10.1007/s10519-005-9003-1.

Luciano, M. *et al.* (2010) ‘Shared genetic aetiology between cognitive ability and cardiovascular disease risk factors: Generation Scotland’s Scottish family health study’, *Intelligence*, 38(3), pp. 304–313. doi: 10.1016/j.intell.2010.03.002.

Luciano, M. *et al.* (2018) ‘Association analysis in over 329 , 000 individuals identifies 116 independent variants influencing neuroticism’, *Nature Genetics*. Nature Publishing Group, 50(January), pp. 6–11. doi: 10.1038/s41588-017-0013-8.

Luo, X. *et al.* (2016) ‘Does refining the phenotype improve replication rates? A review and replication of candidate gene studies on Major Depressive Disorder and Chronic Major Depressive Disorder’, *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics*, 171(2), pp. 215–236. doi: 10.1002/ajmg.b.32396.

Lyall, D. M. *et al.* (2016) ‘Cognitive Test Scores in UK Biobank: Data Reduction in 480,416 Participants and Longitudinal Stability in 20,346 Participants’, *PLOS ONE*. Edited by H. Reddy. Public Library of Science, 11(4), p. e0154222. doi: 10.1371/journal.pone.0154222.

Mackay, T. F. C. (2014) ‘Epistasis and quantitative traits: using model organisms to study gene-gene interactions.’, *Nature reviews. Genetics*, 15(1), pp. 22–33. doi: 10.1038/nrg3627.

Mackay, T. F. and Moore, J. H. (2014) ‘Why epistasis is important for tackling

complex human disease genetics', *Genome Medicine*. BioMed Central, 6(6), p. 125. doi: 10.1186/gm561.

Maeshima, H. *et al.* (2013) 'Time course for memory dysfunction in early-life and late-life major depression: A longitudinal study from the Juntendo university mood disorder project', *Journal of Affective Disorders*, 151(1), pp. 66–70. doi: 10.1016/j.jad.2013.05.050.

Manichaikul, A. *et al.* (2010) 'Robust relationship inference in genome-wide association studies', *Bioinformatics*, 26(22), pp. 2867–2873. doi: 10.1093/bioinformatics/btq559.

Marceau, K. *et al.* (2016) 'The Prenatal Environment in Twin Studies: A Review on Chorionicity', *Behavior Genetics*. Springer, pp. 286–303. doi: 10.1007/s10519-016-9782-6.

Matthews, A. G., Finkelstein, D. M. and Betensky, R. A. (2008) 'Analysis of familial aggregation studies with complex ascertainment schemes', *Statistics in Medicine*. NIH Public Access, 27(24), pp. 5076–5092. doi: 10.1002/sim.3327.

Matthews, G., Deary, I. J. and Whiteman, M. C. (2009) *Personality traits, Third edition, Personality Traits, Third Edition*. doi: 10.1017/CBO9780511812743.

Meijsen, J. J. *et al.* (2018) 'Using tree-based methods for detection of gene-gene interactions in the presence of a polygenic signal: simulation study with application to educational attainment in the Generation Scotland Cohort Study', *Bioinformatics*, 33(17), pp. 2699–2705. doi: 10.1093/bioinformatics/bty462.

Mezuk, B. *et al.* (2008) 'Depression and type 2 diabetes over the lifespan: A meta-analysis', *Diabetes Care*, 31(12), pp. 2383–2390. doi: 10.2337/dc08-0985.

de Moor, M. H. M. *et al.* (2015) 'Meta-analysis of Genome-wide Association Studies for Neuroticism, and the Polygenic Association With Major Depressive Disorder', *JAMA Psychiatry*, 72(7), p. 642. doi: 10.1001/jamapsychiatry.2015.0554.

Muris, P. *et al.* (2005) 'Mediating effects of rumination and worry on the links between neuroticism, anxiety and depression', *Personality and Individual*

Differences, 39(6), pp. 1105–1111. doi: 10.1016/j.paid.2005.04.005.

Murray, C. J. *et al.* (1996) 'Evidence-based health policy--lessons from the Global Burden of Disease Study.', *Science (New York, N.Y.)*, 274(5288), pp. 740–3. doi: 10.1126/SCIENCE.274.5288.740.

Navrady, L. B. *et al.* (2017) 'Intelligence and neuroticism in relation to depression and psychological distress: Evidence from two large population cohorts', *European Psychiatry*. Elsevier Masson, 43, pp. 58–65. doi: 10.1016/j.eurpsy.2016.12.012.

Nelson, H. E. (1982) 'The National Adult Reading Test (NART): Test Manual.', *Windsor, UK: NFER-Nelson*, 124(3), pp. 0–25. doi: Thesis_references-Converted #319.

Nicodemus, K. K. *et al.* (2014) 'Variability in Working Memory Performance Explained by Epistasis vs Polygenic Scores in the ZNF804A Pathway.', *JAMA Psychiatry*, 71, pp. 778–785. doi: 10.1001/jamapsychiatry.2014.528.

Nouwen, A. *et al.* (2010) 'Type 2 diabetes mellitus as a risk factor for the onset of depression: A systematic review and meta-analysis', *Diabetologia*, 53(12), pp. 2480–2486. doi: 10.1007/s00125-010-1874-x.

Nunkesser, R. *et al.* (2007) 'Detecting high-order interactions of single nucleotide polymorphisms using genetic programming', *Bioinformatics*, 23(24), pp. 3280–3288. doi: 10.1093/bioinformatics/btm522.

Nyholt, D. R. (2000) 'All LODs are not created equal.', *American journal of human genetics*. Elsevier, 67(2), pp. 282–288. doi: 10.1086/303029.

Okbay, A. *et al.* (2016) 'Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses', *Nature Genetics*, 48(6), pp. 624–633. doi: 10.1038/ng.3552.

Oliffe, J. L. *et al.* (2016) 'Stigma in Male Depression and Suicide: A Canadian Sex Comparison Study', *Community Mental Health Journal*. Springer, 52(3), pp. 302–310. doi: 10.1007/s10597-015-9986-x.

Oliffe, J. L. and Phillips, M. J. (2008) ‘Men, depression and masculinities: A review and recommendations’, *Journal of Men’s Health*, pp. 194–202. doi: 10.1016/j.jomh.2008.03.016.

Patel, T. and Kurdi, M. S. (2015) ‘A comparative study between oral melatonin and oral midazolam on preoperative anxiety, cognitive, and psychomotor functions.’, *Journal of anaesthesiology, clinical pharmacology*, 31(1), pp. 37–43. doi: 10.4103/0970-9185.150534.

Pilia, G. *et al.* (2006) ‘Heritability of cardiovascular and personality traits in 6,148 Sardinians’, *PLoS Genetics*, 2(8), pp. 1207–1223. doi: 10.1371/journal.pgen.0020132.

Plomin, R. *et al.* (1997) ‘Nature, Nurture, and Cognitive Development from 1 to 16 Years: A Parent-Offspring Adoption Study’, *Psychological Science*. Sage Publications, Inc.Association for Psychological Science, pp. 442–447. doi: 10.2307/40063231.

Plomin, R. *et al.* (2004) ‘A functional polymorphism in the succinate-semialdehyde dehydrogenase (aldehyde dehydrogenase 5 family, member A1) gene is associated with cognitive ability’, *Molecular Psychiatry*. Nature Publishing Group, 9(6), pp. 582–586. doi: 10.1038/sj.mp.4001441.

Plomin, R. and Daniels, D. (2011) ‘Why are children in the same family so different from one another?’, *International Journal of Epidemiology*, 40(3), pp. 563–582. doi: 10.1093/ije/dyq148.

Plomin, R., DeFries, J. C. and McClearn, G. E. (1990) *Behavioral genetics : a primer, A Series of books in psychology*.

Porta, M. S. and International Epidemiological Association. (2008) *A dictionary of epidemiology*. Available at: https://global.oup.com/academic/product/a-dictionary-of-epidemiology-9780199976737?cc=us&lang=en (Accessed: 4 July 2017).

Posthuma, D. *et al.* (2005) ‘A genomewide scan for intelligence identifies quantitative trait loci on 2q and 6p’, *Am J Hum Genet*. Elsevier, 77(2), pp. 318–326.

doi: S0002-9297(07)62921-8 [pii]\n10.1086/432647.

Posthuma, D., De Geus, E. J. C. and Boomsma, D. I. (2001) 'Perceptual speed and IQ are associated through common genetic factors', *Behavior Genetics*. Kluwer Academic Publishers-Plenum Publishers, 31(6), pp. 593–602. doi: 10.1023/A:1013349512683.

Power, R. A. and Pluess, M. (2015) 'Heritability estimates of the Big Five personality traits based on common genetic variants', *Translational Psychiatry*, 5(7). doi: 10.1038/tp.2015.96.

Pulst, S. M. (1999) 'Genetic linkage analysis.', *Archives of neurology*. Johns Hopkins University Press, Baltimore, Md, 56(6), pp. 667–672. doi: 10.1001/archneur.56.6.667.

Quinlan, J. R. (1992) *C4.5: Programs for Machine Learning*, *Morgan Kaufmann San Mateo California*. Morgan Kaufmann Publishers. doi: 10.1016/S0019-9958(62)90649-6.

R Core Team (2017) 'R Core Team (2017). R: A language and environment for statistical computing.', *R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.*, p. R Foundation for Statistical Computing.

Radstake, T. R. D. J. *et al.* (2010) 'Genome-wide association study of systemic sclerosis identifies CD247 as a new susceptibility locus.', *Nature genetics*. NIH Public Access, 42(5), pp. 426–9. doi: 10.1038/ng.565.

Raven, J. C., Raven, J. and Court, J. H. (1988) *The Mill Hill vocabulary scale, Manual for Raven's progressive matrices and vocabulary scales*.

Ripke, S. *et al.* (2013) 'A mega-analysis of genome-wide association studies for major depressive disorder', *Molecular Psychiatry*, 18(4), pp. 497–511. doi: 10.1038/mp.2012.21.

Ritchie, S. J. (2015) *Intelligence : all that matters*.

Roelofs, J. *et al.* (2008) 'Rumination and worrying as possible mediators in the

relation between neuroticism and symptoms of depression and anxiety in clinically depressed individuals', *Behaviour Research and Therapy*, 46(12), pp. 1283–1289. doi: 10.1016/j.brat.2008.10.002.

Ruczinski, I., Kooperberg, C. and LeBlanc, M. (2003) 'Logic Regression', *Journal of Computational and Graphical Statistics*, 12(3), pp. 475–511. doi: 10.1198/1061860032238.

Ruczinski, I., Kooperberg, C. and LeBlanc, M. L. (2004) 'Exploring interactions in high-dimensional genomic data: An overview of Logic Regression, with applications', *Journal of Multivariate Analysis*, pp. 178–195. doi: 10.1016/j.jmva.2004.02.010.

Sackton, T. B. and Hartl, D. L. (2016) 'Genotypic Context and Epistasis in Individuals and Populations', *Cell*, pp. 279–287. doi: 10.1016/j.cell.2016.06.047.

Salthouse, T. A. (1996) 'The processing-speed theory of adult age differences in cognition.', *Psychological Review*, 103(3), pp. 403–428. doi: 10.1037/0033-295X.103.3.403.

Salzberg, S. L. (1994) 'C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993', *Machine Learning*. Kluwer Academic Publishers, 16(3), pp. 235–240. doi: 10.1007/BF00993309.

Schwender, H. (2007) 'Statistical Analysis of Genotype and Gene Expression Data Dissertation'.

Small, B. J. *et al.* (2004) 'Apolipoprotein E and Cognitive Performance: A Meta-Analysis.', *Psychology and Aging*, 19(4), pp. 592–600. doi: 10.1037/0882-7974.19.4.592.

Smith, B. H. *et al.* (2006) 'Generation Scotland: the Scottish Family Health Study; a new resource for researching genes and heritability.', *BMC medical genetics*, 7(1), p. 74. doi: 10.1186/1471-2350-7-74.

Smith, B. H. *et al.* (2013) 'Cohort profile: Generation scotland: Scottish family health study (GS: SFHS). The study, its participants and their potential for genetic

research on health and illness', *International Journal of Epidemiology*, 42(3), pp. 689–700. doi: 10.1093/ije/dys084.

Smith, D. J. *et al.* (2013) 'Prevalence and characteristics of probable major depression and bipolar disorder within UK Biobank: Cross-sectional study of 172,751 participants', *PLoS ONE*. Edited by J. B. Potash. Public Library of Science, 8(11), p. e75362. doi: 10.1371/journal.pone.0075362.

Smith, D. J. *et al.* (2016) 'Genome-wide analysis of over 106 000 individuals identifies 9 neuroticism-associated loci', *Molecular Psychiatry*, 21(6), pp. 749–757. doi: 10.1038/mp.2016.49.

Sniekers, S. *et al.* (2017) 'Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence', *Nature Publishing Group*. Nature Research, 49(7), pp. 1107–1112. doi: 10.1038/ng.3869.

Snyder, H. R. (2013) 'Major depressive disorder is associated with broad impairments on neuropsychological measures of executive function: A meta-analysis and review.', *Psychological Bulletin*. NIH Public Access, 139(1), pp. 81–132. doi: 10.1037/a0028727.

So, H. C., Li, M. and Sham, P. C. (2011) 'Uncovering the total heritability explained by all true susceptibility variants in a genome-wide association study', *Genetic Epidemiology*, 35(6), pp. 447–456. doi: 10.1002/gepi.20593.

Sudlow, C. *et al.* (2015) 'UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age', *PLoS Medicine*, 12(3), p. e1001779. doi: 10.1371/journal.pmed.1001779.

Sullivan, P. F., Neale, M. C. and Kendler, K. S. (2000) 'Genetic epidemiology of major depression: Review and meta-analysis', *American Journal of Psychiatry*, pp. 1552–1562. doi: 10.1176/appi.ajp.157.10.1552.

Tabor, H. K., Risch, N. J. and Myers, R. M. (2002) 'Candidate-gene approaches for studying complex genetic traits: Practical considerations.', *Nature Reviews. Genetics*. Nature Publishing Group, 3(5), pp. 391–397. doi: 10.1038/nrg796.

Talarowska, M., Zajaczkowska, M. and Galecki, P. (2015) ‘Cognitive functions in first-episode depression and recurrent depressive disorder’, *Psychiatria Danubina*, 27(1), pp. 38–43.

Thomas, C. M. and Morris, S. (2003) ‘Cost of depression among adults in England in 2000’, *British Journal of Psychiatry*, 183(DEC.), pp. 514–519. doi: 10.1192/bjp.183.6.514.

Tibshirani, R. (1996) ‘Regression Selection and Shrinkage via the Lasso’, *Journal of the Royal Statistical Society B*, pp. 267–288. doi: 10.2307/2346178.

Travis, S. *et al.* (2014) ‘Dentate gyrus volume and memory performance in major depressive disorder.’, *Journal of affective disorders*, 172C, pp. 159–164. doi: 10.1016/j.jad.2014.09.048.

Tucker-Drob, E. M. (2009) ‘Differentiation of cognitive abilities across the life span.’, *Developmental Psychology*, 45(4), pp. 1097–1118. doi: 10.1037/a0015864.

Vukasović, T. and Bratko, D. (2015) ‘Heritability of personality: A meta-analysis of behavior genetic studies’, *Psychological Bulletin*, 141(4), pp. 769–785. doi: 10.1037/bul0000017.

Walters, R., Laurin, C. and Lubke, G. H. (2012) ‘An integrated approach to reduce the impact of minor allele frequency and linkage disequilibrium on variable importance measures for genome-wide data’, *Bioinformatics*. Oxford University Press, 28(20), pp. 2615–2623. doi: 10.1093/bioinformatics/bts483.

Webber, C. (2017) ‘Epistasis in Neuropsychiatric Disorders’, *Trends in Genetics*, pp. 256–265. doi: 10.1016/j.tig.2017.01.009.

Wechsler D. (1998) ‘WAIS-III UK Wechsler Adult Intelligence Scale.’, *Psychological Corporation*.

Winterer, G. and Goldman, D. (2003) ‘Genetics of human prefrontal function’, *Brain Research Reviews*, pp. 134–163. doi: 10.1016/S0165-0173(03)00205-4.

Witte, J. S. (2010) ‘Genome-Wide Association Studies and Beyond’, *Annual Review*

of Public Health, 31(1), pp. 9–20. doi: 10.1146/annurev.publhealth.012809.103723.

World Health Organisation (2004) 'Rules and guidelines for mortality and morbidity coding', *International Classification of Diseases and Related Health Problems. Tenth Revision. Volume 2*, 92(3), pp. 31–92. doi: 10.1016/j.jclinepi.2009.09.002.

Wray, N. R. and Sullivan, P. F. (2018) 'Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression', *Nature Genetics*. doi: 10.1101/167577.

Wu, X. *et al.* (2008) 'Top 10 algorithms in data mining', *Knowledge and Information Systems*, 14(1), pp. 1–37. doi: 10.1007/s10115-007-0114-2.

Yang, J. *et al.* (2010) 'Common SNPs explain a large proportion of the heritability for human heigth', *Nature genetics*, 42(7), pp. 565–569. doi: 10.1038/ng.608.

Yang, J. *et al.* (2011) 'GCTA: a tool for genome-wide complex trait analysis.', *American journal of human genetics*, 88(1), pp. 76–82. doi: 10.1016/j.ajhg.2010.11.011.

Zhang, W. *et al.* (2015) 'Structural basis of arc binding to synaptic proteins: implications for cognitive disease.', *Neuron*, 86(2), pp. 490–500. doi: 10.1016/j.neuron.2015.03.030.

Zhao, Y. *et al.* (2012) 'Correction for population stratification in random forest analysis', *International Journal of Epidemiology*. Oxford University Press, 41(6), pp. 1798–1806. doi: 10.1093/ije/dys183.

Zisook, S. *et al.* (2007) 'Effect of age at onset on the course of major depressive disorder', *American Journal of Psychiatry*, 164(10), pp. 1539–1546. doi: 10.1176/appi.ajp.2007.06101757.

Zou, H. and Hastie, T. (2005) 'Regularization and variable selection via the elastic net (vol B 67, pg 301, 2005)', *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 67(2), p. 768. doi: DOI 10.1111/j.1467-9868.2005.00527.x.

Zubenko, G. S. *et al.* (2003) 'Genome-wide linkage survey for genetic loci that

influence the development of depressive disorders in families with recurrent, early-onset, major depression', *American Journal of Medical Genetics*, 123B(1), pp. 1–18. doi: 10.1002/ajmg.b.20073.

7 Supplementary material

This required binary responses (1-yes; 0 –no) to the following items:
1) Does your mood often go up and down?
2) Do you ever feel 'just miserable' for no reason?
3) Are you an irritable person?
4) Are your feelings easily hurt?
5) Do you often feel 'fed-up'?
6) Would you call yourself a nervous person?
7) Are you a worrier?
8) Would you call yourself tense or 'highly strung'?
9) Do you worry too long after an embarrassing experience?
10) Do you suffer from 'nerves'?
11) Do you often feel lonely?
12) Are you often troubled by feelings of guilt?

Supplementary Table: 4.1: 12-item Eysenck Personality Questionnaire-Revised (EPQ-R) Short Form

Ruleset number	Ruleset	Outcome	N individuals
1	rs2371924_G in [1-2] rs1526335_A = 0 rs12904777_C = 2	-1.1655182	313
2	rs2371924_G in [1-2] rs610789_G in [1-2] rs2750007_A = 0 rs299175_A = 0	-1.3004745	805
3	rs884718_G = 0 rs1526335_A = 0 rs2750007_A = 0 rs299175_A = 0	-1.4905023	670
4	rs2371924_G in [1-2] rs426029_A in [0-1] rs10245350_G in [1-2] rs748315_A in [1-2] rs299175_A in [1-2]	-1.0845443	257
5	rs2371924_G in [1-2] rs1526335_A in [1-2]	-0.9212739	1705
6	rs2371924_G in [1-2] rs11496038_G in [0-1] rs10245350_G = 0 rs17277543_A in [0-1] rs299175_A in [1-2]	-0.8686141	2509
7	rs11928265_G = 0 rs11496038_G in [0-1] rs17277543_A = 2 rs12904777_C in [0-1]	-0.3542232	951
8	rs2371924_G in [1-2] rs11928265_G in [1-2] rs10245350_G = 0 rs7977649_A in [1-2] rs17277543_A = 2 rs12904777_C in [0-1]	-0.057553	95
9	rs2371924_G = 0	-0.435197	1938
10	rs1526335_A = 0 rs12904777_C in [0-1] rs2750007_A in [1-2] rs299175_A = 0	0.2406738	187
11	rs2371924_G in [1-2] rs884718_G in [1-2]	0.753508	244

	rs610789_G = 0 rs2750007_A = 0 rs299175_A = 0		
12	rs426029_A = 2 rs2527506_A = 0 rs11496038_G in [0-1] rs10245350_G in [1-2] rs1526335_A = 0 rs299175_A in [1-2]	1.6455431	35
13	rs2371924_G in [1-2] rs11928265_G in [1-2] rs10245350_G = 0 rs4298432_G = 0 rs1526335_A = 0 rs7977649_A = 0 rs17277543_A = 2 rs12904777_C in [0-1] rs299175_A in [1-2]	1.1856023	34
14	rs2371924_G in [1-2] rs11496038_G = 2 rs1526335_A = 0 rs12904777_C in [0-1] rs299175_A in [1-2]	1.1856023	133
15	rs2371924_G in [1-2] rs426029_A in [0-1] rs11496038_G in [0-1] rs10245350_G in [1-2] rs1526335_A = 0 rs748315_A = 0 rs299175_A in [1-2]	1.8347157	81
16	rs2371924_G in [1-2] rs11928265_G in [1-2] rs10245350_G = 0 rs4298432_G in [1-2] rs1526335_A = 0 rs7977649_A = 0 rs17277543_A = 2 rs299175_A in [1-2]	5.3216478	40
17	rs2371924_G in [1-2] rs426029_A = 2 rs2527506_A in [1-2] rs11496038_G in [0-1]	6.1586112	30

	rs10245350_G in [1-2] rs1526335_A = 0 rs12904777_C in [0-1]		

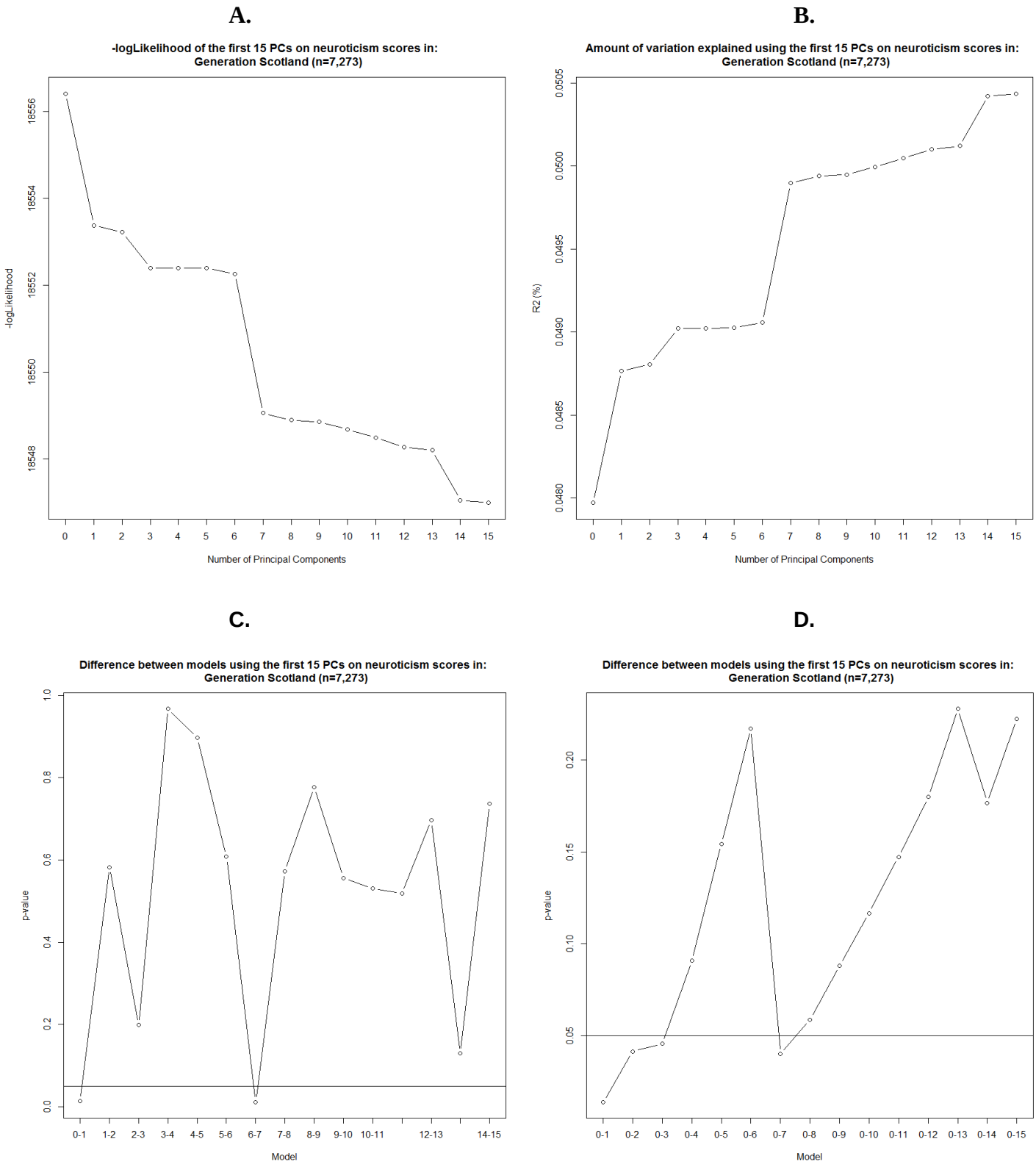
Supplementary Table: 4.2:17 C5.0 rules observed associated with neuroticism in Generation Scotland. Values reported in square brackets represent the genotypes i.e. SNP_x [1-2] = genotype 1 or 2 for SNP_x.

Component	Coefficient	Component	Coefficient
Rule2	-0.3315	rs13317333_A	-0.0171
Rule3	-0.241	rs4769066_G	-0.017
Rule1	-0.2153	rs918949_G	-0.0166
rs3753183_C	-0.0889	rs17564279_G	-0.0165
rs17104742_A	-0.0872	rs16849142_A	-0.0156
Rule5	-0.0868	rs12468177_A	-0.0152
rs7866089_A	-0.0619	rs722885_G	-0.0141
rs3768405_A	-0.0607	rs12719482_G	-0.0136
Rule6	-0.0471	rs270309_G	-0.0136
rs879143_A	-0.0464	rs1478366_G	-0.0121
rs7722634_A	-0.0439	rs1671150_A	-0.0119
rs13160599_G	-0.0434	rs3932174_A	-0.0111
rs2991010_A	-0.0432	rs867389_G	-0.0108
rs7507140_G	-0.0395	rs428570_C	-0.0107
rs9827448_A	-0.0371	rs2370794_G	-0.0098
rs11198856_A	-0.036	rs10202624_A	-0.0088
rs2235698_G	-0.0339	rs10084852_G	-0.0087
rs1601875_A	-0.0317	rs235777_A	-0.0087
rs9886142_G	-0.0302	rs12827688_A	-0.0071
rs768784_A	-0.0301	rs1954610_G	-0.0068
rs169712_G	-0.03	rs1029409_G	-0.0063
rs4762210_G	-0.0286	rs11059014_A	-0.006
rs11578964_A	-0.0282	rs869834_G	-0.0058
rs4411458_G	-0.0275	rs11665404_C	-0.0055
rs866995_G	-0.0274	rs704017_A	-0.0052
rs221308_G	-0.0263	rs12118937_A	-0.005
rs17479963_A	-0.0258	rs7413698_G	-0.005
rs4661142_A	-0.0251	rs9920603_A	-0.0048
rs2387326_A	-0.0241	rs16891104_G	-0.0046
rs12033709_A	-0.0235	rs10036822_G	-0.0042
rs2800_G	-0.0232	rs12051618_A	-0.004
rs12151513_C	-0.0218	rs12638703_A	-0.0033
rs8665_G	-0.0211	rs10016747_A	-0.0033
rs11189833_A	-0.0206	rs10221449_C	-0.0032
rs231880_A	-0.0199	rs9603687_A	-0.0029
rs4633802_G	-0.0198	rs9523762_A	-0.0028
rs12909221_G	-0.0198	rs7818637_G	-0.0027
rs823009_A	-0.0194	rs301694_A	-0.0027
rs2406706_A	-0.0191	rs7722079_C	-0.0016
rs9376740_G	-0.0188	rs7968211_G	-0.001

rs2338967_C	-0.0003	rs2161765_G	0.0113
rs10945513_A	-0.0002	rs1128349_A	0.0116
rs12477044_A	-8.01E-06	rs2853418_A	0.0117
rs9357429_A	0.0002	rs9445284_A	0.0122
rs8002697_A	0.0002	rs553780_A	0.0126
rs4778006_G	0.0003	rs12944785_G	0.0136
rs11207811_C	0.0004	rs11706338_C	0.0143
rs17351628_A	0.0007	rs17793937_G	0.0147
rs4280764_G	0.0014	rs11930915_A	0.0148
rs10814288_A	0.0014	rs243021_A	0.0158
rs9367018_G	0.0015	rs719802_A	0.0168
rs6798084_G	0.0018	rs4742447_G	0.0174
rs11872992_A	0.0019	rs7835387_G	0.0185
rs2930313_G	0.002	rs352024_A	0.0193
rs2802984_A	0.0025	rs2215290_A	0.0195
rs6545946_A	0.0027	rs600671_A	0.0212
rs77905_G	0.0028	rs6553050_G	0.0224
rs1051858_G	0.0029	rs987771_G	0.0232
rs1170665_G	0.0031	rs4795856_A	0.0236
rs4581716_G	0.0035	rs11690032_A	0.0266
rs11119014_G	0.004	rs2830634_A	0.0285
rs10976131_G	0.0045	rs11205387_A	0.0288
rs10996914_G	0.0046	rs6452194_G	0.031
rs12895581_G	0.005	rs12233479_A	0.0342
rs7240537_A	0.0052	rs7813088_A	0.0355
rs2521259_A	0.0053	rs7656865_G	0.0355
rs7776080_A	0.0071	rs12611768_G	0.0364
rs7104786_A	0.0072	rs877707_G	0.0437
rs6939316_A	0.0076	rs2425024_C	0.044
rs10828753_A	0.0079	rs7041938_C	0.0458
rs344108_A	0.0082	rs1615246_G	0.0475
rs2179176_A	0.0084	rs6470016_G	0.0488
rs34575650_A	0.0092	rs6785153_C	0.0502
rs4776010_A	0.0092	Rule11	0.054
rs1126230_C	0.0098	rs12770361_A	0.0589
rs934668_G	0.0103	rs17463085_A	0.0681
rs4524755_A	0.0108	rs17157128_G	0.0703
rs4970856_A	0.011	Rule9	0.0721
rs2419778_A	0.011	Rule15	0.6252
rs2474254_G	0.0113	Rule14	0.6386
Rule16	2.7133		

Rule17	2.9368
Rule16	2.7133
Rule17	2.9368

Supplementary Table 4.3: Coefficients of all 162 variables selected by LASSO. Coefficients in red are <0 while green coefficient are >0.



Supplementary Figures 4.1A-D: Investigating population substructure in Generation Scotland by calculating for the first 15 PCs: *A) The –logLikelihood, B) R^2, C) difference in fitness* including PCs and D) difference in fitness* compared to base model. *fitness is defined as a measure of how well the created model fits a set of observations.*